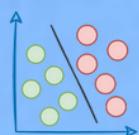
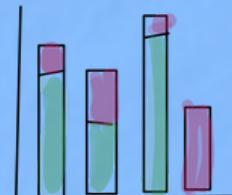
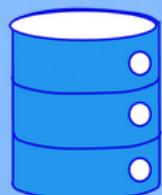
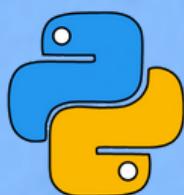


**FREE**

# DATA SCIENCE

**FULL ARCHIVE**

200+ Python & Data  
Science Tips



Daily Dose of  
Data Science



[avichawla.substack.com](https://avichawla.substack.com)



## Table of Contents

<i>Breathing KMeans: A Better and Faster Alternative to KMeans .....</i>	8
<i>How Many Dimensions Should You Reduce Your Data To When Using PCA?.....</i>	11
<i>🚀 Mito Just Got Supercharged With AI!.....</i>	14
<i>Be Cautious Before Drawing Any Conclusions Using Summary Statistics .....</i>	16
<i>Use Custom Python Objects In A Boolean Context.....</i>	18
<i>A Visual Guide To Sampling Techniques in Machine Learning.....</i>	20
<i>You Were Probably Given Incomplete Info About A Tuple's Immutability .....</i>	24
<i>A Simple Trick That Significantly Improves The Quality of Matplotlib Plots .....</i>	26
<i>A Visual and Overly Simplified Guide to PCA.....</i>	28
<i>Supercharge Your Jupyter Kernel With ipyflow .....</i>	31
<i>A Lesser-known Feature of Creating Plots with Plotly .....</i>	33
<i>The Limitation Of Euclidean Distance Which Many Often Ignore.....</i>	35
<i>Visualising The Impact Of Regularisation Parameter .....</i>	38
<i>AutoProfiler: Automatically Profile Your DataFrame As You Work.....</i>	40
<i>A Little Bit Of Extra Effort Can Hugely Transform Your Storytelling Skills .....</i>	42
<i>A Nasty Hidden Feature of Python That Many Programmers Aren't Aware Of.</i>	44
<i>Interactively Visualise A Decision Tree With A Sankey Diagram .....</i>	47
<i>Use Histograms With Caution. They Are Highly Misleading! .....</i>	49
<i>Three Simple Ways To (Instantly) Make Your Scatter Plots Clutter Free .....</i>	51
<i>A (Highly) Important Point to Consider Before You Use KMeans Next Time .....</i>	54
<i>Why You Should Avoid Appending Rows To A DataFrame .....</i>	57
<i>Matplotlib Has Numerous Hidden Gems. Here's One of Them.....</i>	59
<i>A Counterintuitive Thing About Python Dictionaries .....</i>	61
<i>Probably The Fastest Way To Execute Your Python Code .....</i>	64
<i>Are You Sure You Are Using The Correct Pandas Terminologies? .....</i>	66
<i>Is Class Imbalance Always A Big Problem To Deal With?.....</i>	69
<i>A Simple Trick That Will Make Heatmaps More Elegant .....</i>	71
<i>A Visual Comparison Between Locality and Density-based Clustering .....</i>	73
<i>Why Don't We Call It Logistic Classification Instead? .....</i>	74
<i>A Typical Thing About Decision Trees Which Many Often Ignore.....</i>	76



<i>Always Validate Your Output Variable Before Using Linear Regression .....</i>	77
<i>A Counterintuitive Fact About Python Functions .....</i>	78
<i>Why Is It Important To Shuffle Your Dataset Before Training An ML Model .....</i>	79
<i>The Limitations Of Heatmap That Are Slowing Down Your Data Analysis.....</i>	80
<i>The Limitation Of Pearson Correlation Which Many Often Ignore .....</i>	81
<i>Why Are We Typically Advised To Set Seeds for Random Generators?.....</i>	82
<i>An Underrated Technique To Improve Your Data Visualizations .....</i>	83
<i>A No-Code Tool to Create Charts and Pivot Tables in Jupyter.....</i>	84
<i>If You Are Not Able To Code A Vectorized Approach, Try This. ....</i>	85
<i>Why Are We Typically Advised To Never Iterate Over A DataFrame?.....</i>	87
<i>Manipulating Mutable Objects In Python Can Get Confusing At Times .....</i>	88
<i>This Small Tweak Can Significantly Boost The Run-time of KMeans .....</i>	90
<i>Most Python Programmers Don't Know This About Python OOP .....</i>	92
<i>Who Said Matplotlib Cannot Create Interactive Plots? .....</i>	94
<i>Don't Create Messy Bar Plots. Instead, Try Bubble Charts!.....</i>	95
<i>You Can Add a List As a Dictionary's Key (Technically)!.....</i>	96
<i>Most ML Folks Often Neglect This While Using Linear Regression .....</i>	97
<i>35 Hidden Python Libraries That Are Absolute Gems .....</i>	98
<i>Use Box Plots With Caution! They May Be Misleading. ....</i>	99
<i>An Underrated Technique To Create Better Data Plots .....</i>	100
<i>The Pandas DataFrame Extension Every Data Scientist Has Been Waiting For</i>	101
<i>Supercharge Shell With Python Using Xonsh .....</i>	102
<i>Most Command-line Users Don't Know This Cool Trick About Using Terminals .....</i>	103
<i>A Simple Trick to Make The Most Out of Pivot Tables in Pandas.....</i>	104
<i>Why Python Does Not Offer True OOP Encapsulation.....</i>	105
<i>Never Worry About Parsing Errors Again While Reading CSV with Pandas.....</i>	106
<i>An Interesting and Lesser-Known Way To Create Plots Using Pandas.....</i>	107
<i>Most Python Programmers Don't Know This About Python For-loops .....</i>	108
<i>How To Enable Function Overloading In Python.....</i>	109
<i>Generate Helpful Hints As You Write Your Pandas Code .....</i>	110
<i>Speedup NumPy Methods 25x With Bottleneck .....</i>	111
<i>Visualizing The Data Transformation of a Neural Network .....</i>	112



<i>Never Refactor Your Code Manually Again. Instead, Use Sourcery!</i> .....	113
<i>Draw The Data You Are Looking For In Seconds</i> .....	114
<i>Style Matplotlib Plots To Make Them More Attractive</i> .....	115
<i>Speed-up Parquet I/O of Pandas by 5x</i> .....	116
<i>40 Open-Source Tools to Supercharge Your Pandas Workflow</i> .....	117
<i>Stop Using The Describe Method in Pandas. Instead, use Skimpy.</i> .....	118
<i>The Right Way to Roll Out Library Updates in Python</i> .....	119
<i>Simple One-Liners to Preview a Decision Tree Using Sklearn</i> .....	120
<i>Stop Using The Describe Method in Pandas. Instead, use Summarytools.</i> .....	121
<i>Never Search Jupyter Notebooks Manually Again To Find Your Code</i> .....	122
<i>F-strings Are Much More Versatile Than You Think</i> .....	123
<i>Is This The Best Animated Guide To KMeans Ever?</i> .....	124
<i>An Effective Yet Underrated Technique To Improve Model Performance</i> .....	125
<i>Create Data Plots Right From The Terminal</i> .....	126
<i>Make Your Matplotlib Plots More Professional</i> .....	127
<i>37 Hidden Python Libraries That Are Absolute Gems</i> .....	128
<i>Preview Your README File Locally In GitHub Style</i> .....	129
<i>Pandas and NumPy Return Different Values for Standard Deviation. Why? ...</i>	130
<i>Visualize Commit History of Git Repo With Beautiful Animations</i> .....	131
<i>Perfplot: Measure, Visualize and Compare Run-time With Ease</i> .....	132
<i>This GUI Tool Can Possibly Save You Hours Of Manual Work</i> .....	133
<i>How Would You Identify Fuzzy Duplicates In A Data With Million Records?....</i>	134
<i>Stop Previewing Raw DataFrames. Instead, Use DataTables.</i> .....	136
<i>A Single Line That Will Make Your Python Code Faster</i> .....	137
<i>Prettify Word Clouds In Python</i> .....	138
<i>How to Encode Categorical Features With Many Categories?</i> .....	139
<i>Calendar Map As A Richer Alternative to Line Plot</i> .....	140
<i>10 Automated EDA Tools That Will Save You Hours Of (Tedious) Work</i> .....	141
<i>Why KMeans May Not Be The Apt Clustering Algorithm Always</i> .....	142
<i>Converting Python To LaTeX Has Possibly Never Been So Simple</i> .....	143
<i>Density Plot As A Richer Alternative to Scatter Plot</i> .....	144
<i>30 Python Libraries to (Hugely) Boost Your Data Science Productivity</i> .....	145
<i>Sklearn One-liner to Generate Synthetic Data</i> .....	146



<i>Label Your Data With The Click Of A Button</i> .....	147
<i>Analyze A Pandas DataFrame Without Code</i> .....	148
<i>Python One-Liner To Create Sketchy Hand-drawn Plots</i> .....	149
<i>70x Faster Pandas By Changing Just One Line of Code</i> .....	150
<i>An Interactive Guide To Master Pandas In One Go</i> .....	151
<i>Make Dot Notation More Powerful in Python</i> .....	152
<i>The Coolest Jupyter Notebook Hack</i> .....	153
<i>Create a Moving Bubbles Chart in Python</i> .....	154
<i>Skorch: Use Scikit-learn API on PyTorch Models</i> .....	155
<i>Reduce Memory Usage Of A Pandas DataFrame By 90%</i> .....	156
<i>An Elegant Way To Perform Shutdown Tasks in Python</i> .....	157
<i>Visualizing Google Search Trends of 2022 using Python</i> .....	158
<i>Create A Racing Bar Chart In Python</i> .....	159
<i>Speed-up Pandas Apply 5x with NumPy</i> .....	160
<i>A No-Code Online Tool To Explore and Understand Neural Networks</i> .....	161
<i>What Are Class Methods and When To Use Them?</i> .....	162
<i>Make Sklearn KMeans 20x times faster</i> .....	163
<i>Speed-up NumPy 20x with Numexpr</i> .....	164
<i>A Lesser-Known Feature of Apply Method In Pandas</i> .....	165
<i>An Elegant Way To Perform Matrix Multiplication</i> .....	166
<i>Create Pandas DataFrame from Dataclass</i> .....	167
<i>Hide Attributes While Printing A Dataclass Object</i> .....	168
<i>List : Tuple :: Set : ?</i> .....	169
<i>Difference Between Dot and Matmul in NumPy</i> .....	170
<i>Run SQL in Jupyter To Analyze A Pandas DataFrame</i> .....	171
<i>Automated Code Refactoring With Sourcery</i> .....	172
<i>__Post_init__: Add Attributes To A Dataclass Object Post Initialization</i> .....	173
<i>Simplify Your Functions With Partial Functions</i> .....	174
<i>When You Should Not Use the head() Method In Pandas</i> .....	175
<i>DotMap: A Better Alternative to Python Dictionary</i> .....	176
<i>Prevent Wild Imports With __all__ in Python</i> .....	177
<i>Three Lesser-known Tips For Reading a CSV File Using Pandas</i> .....	178
<i>The Best File Format To Store A Pandas DataFrame</i> .....	179



<i>Debugging Made Easy With PySnooper</i> .....	180
<i>Lesser-Known Feature of the Merge Method in Pandas</i> .....	181
<i>The Best Way to Use Apply() in Pandas</i> .....	182
<i>Deep Learning Network Debugging Made Easy</i> .....	183
<i>Don't Print NumPy Arrays! Use Lovely-NumPy Instead.</i> .....	184
<i>Performance Comparison of Python 3.11 and Python 3.10</i> .....	185
<i>View Documentation in Jupyter Notebook</i> .....	186
<i>A No-code Tool To Understand Your Data Quickly</i> .....	187
<i>Why 256 is 256 But 257 is not 257?</i> .....	188
<i>Make a Class Object Behave Like a Function</i> .....	190
<i>Lesser-known feature of Pickle Files</i> .....	192
<i>Dot Plot: A Potential Alternative to Bar Plot</i> .....	194
<i>Why Correlation (and Other Statistics) Can Be Misleading</i> .....	195
<i>Supercharge value_counts() Method in Pandas With Sideload</i> .....	196
<i>Write Your Own Flavor Of Pandas</i> .....	197
<i>CodeSquire: The AI Coding Assistant You Should Use Over GitHub Copilot</i> .....	198
<i>Vectorization Does Not Always Guarantee Better Performance</i> .....	199
<i>In Defense of Match-case Statements in Python</i> .....	200
<i>Enrich Your Notebook With Interactive Controls</i> .....	202
<i>Get Notified When Jupyter Cell Has Executed</i> .....	204
<i>Data Analysis Using No-Code Pandas In Jupyter</i> .....	205
<i>Using Dictionaries In Place of If-conditions</i> .....	206
<i>Clear Cell Output In Jupyter Notebook During Run-time</i> .....	208
<i>A Hidden Feature of Describe Method In Pandas</i> .....	209
<i>Use Slotted Class To Improve Your Python Code</i> .....	210
<i>Stop Analysing Raw Tables. Use Styling Instead!</i> .....	211
<i>Explore CSV Data Right From The Terminal</i> .....	212
<i>Generate Your Own Fake Data In Seconds</i> .....	213
<i>Import Your Python Package as a Module</i> .....	214
<i>Specify Loops and Runs In %%timeit</i> .....	215
<i>Waterfall Charts: A Better Alternative to Line/Bar Plot</i> .....	216
<i>Hexbin Plots As A Richer Alternative to Scatter Plots</i> .....	217
<i>Importing Modules Made Easy with Pyforest</i> .....	218



<i>Analyse Flow Data With Sankey Diagrams</i> .....	220
<i>Feature Tracking Made Simple In Sklearn Transformers</i> .....	222
<i>Lesser-known Feature of f-strings in Python</i> .....	224
<i>Don't Use time.time() To Measure Execution Time</i> .....	225
<i>Now You Can Use DALL-E With OpenAI API</i> .....	226
<i>Polynomial Linear Regression Plot Made Easy With Seaborn</i> .....	227
<i>Retrieve Previously Computed Output In Jupyter Notebook</i> .....	228
<i>Parallelize Pandas Apply() With Swifter</i> .....	229
<i>Create DataFrame Hassle-free By Using Clipboard</i> .....	230
<i>Run Python Project Directory As A Script</i> .....	231
<i>Inspect Program Flow with IceCream</i> .....	232
<i>Don't Create Conditional Columns in Pandas with Apply</i> .....	233
<i>Pretty Plotting With Pandas</i> .....	234
<i>Build Baseline Models Effortlessly With Sklearn</i> .....	235
<i>Fine-grained Error Tracking With Python 3.11</i> .....	236
<i>Find Your Code Hiding In Some Jupyter Notebook With Ease</i> .....	237
<i>Restart the Kernel Without Losing Variables</i> .....	238
<i>How to Read Multiple CSV Files Efficiently</i> .....	239
<i>Elegantly Plot the Decision Boundary of a Classifier</i> .....	241
<i>An Elegant Way to Import Metrics From Sklearn</i> .....	242
<i>Configure Sklearn To Output Pandas DataFrame</i> .....	243
<i>Display Progress Bar With Apply() in Pandas</i> .....	244
<i>Modify a Function During Run-time</i> .....	245
<i>Regression Plot Made Easy with Plotly</i> .....	246
<i>Polynomial Linear Regression with NumPy</i> .....	247
<i>Alter the Datatype of Multiple Columns at Once</i> .....	248
<i>Datatype For Handling Missing Valued Columns in Pandas</i> .....	249
<i>Parallelize Pandas with Pandarallel</i> .....	250
<i>Why you should not dump DataFrames to a CSV</i> .....	251
<i>Save Memory with Python Generators</i> .....	253
<i>Don't use print() to debug your code.</i> .....	254
<i>Find Unused Python Code With Ease</i> .....	256
<i>Define the Correct DataType for Categorical Columns</i> .....	257

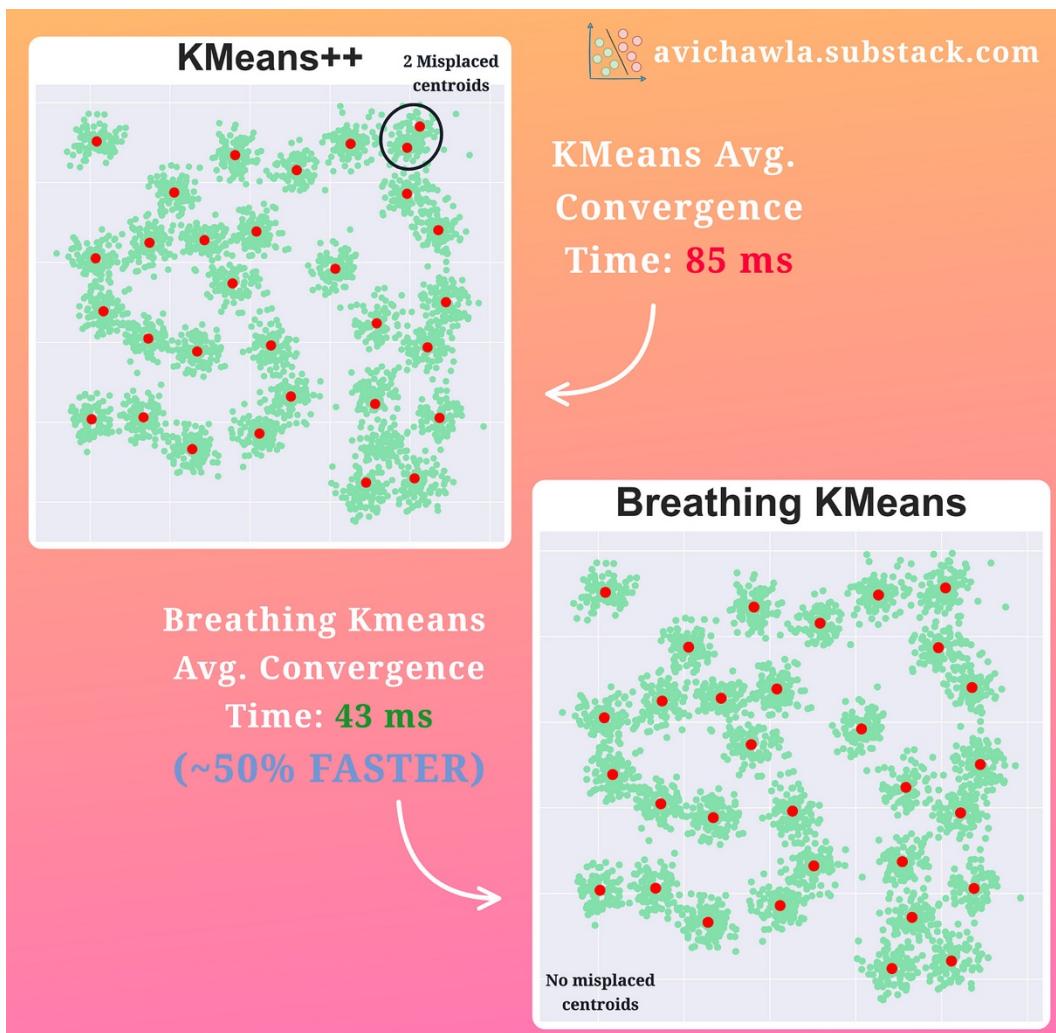


avichawla.substack.com

<b><i>Transfer Variables Between Jupyter Notebooks</i></b> .....	<b>258</b>
<b><i>Why You Should Not Read CSVs with Pandas</i></b> .....	<b>259</b>
<b><i>Modify Python Code During Run-Time</i></b> .....	<b>260</b>
<b><i>Handle Missing Data With Missingno</i></b> .....	<b>261</b>



# Breathing KMeans: A Better and Faster Alternative to KMeans



The performance of KMeans is entirely dependent on the centroid initialization step. Thus, obtaining inaccurate clusters is highly likely.

While KMeans++ offers smarter centroid initialization, it does not always guarantee accurate convergence (read how KMeans++ works in my [previous post](#)). This is especially true when the number of clusters is high. Here, repeating the algorithm may help. But it introduces an unnecessary overhead in run-time.

Instead, Breathing KMeans is a better alternative here. Here's how it works:

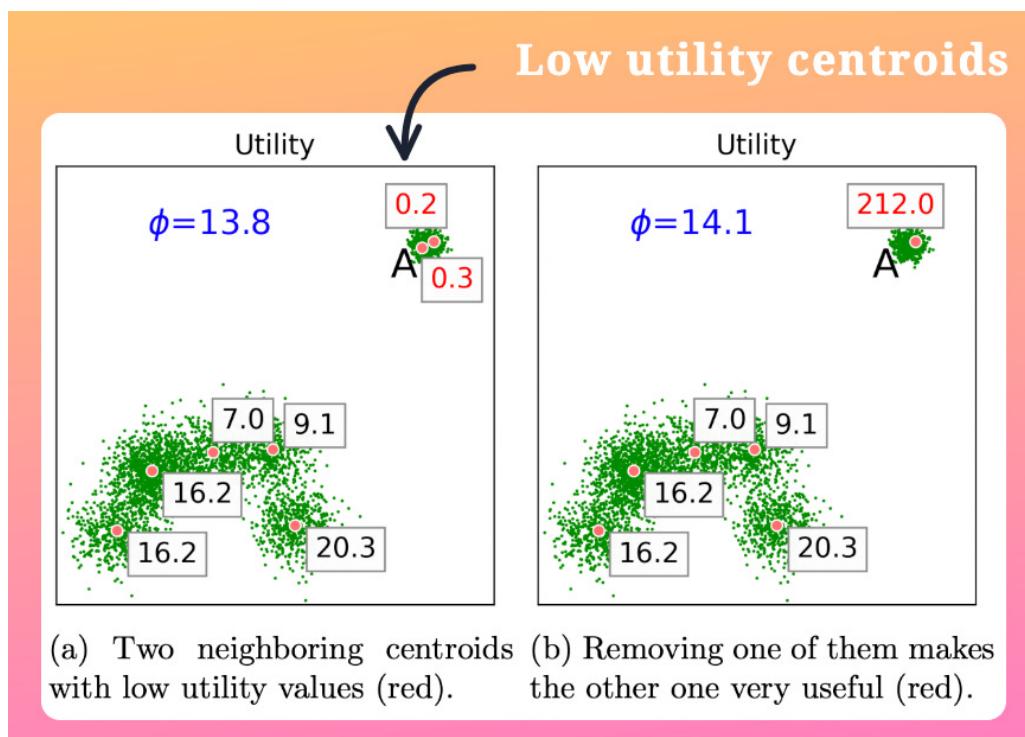
- **Step 1:** Initialise  $k$  centroids and run KMeans without repeating. In other words, don't re-run it with different initializations. Just run it once.



- **Step 2 — Breathe in step:** Add  $m$  new centroids and run KMeans with  $(k+m)$  centroids without repeating.
- **Step 3 — Breathe out step:** Remove  $m$  centroids from existing  $(k+m)$  centroids. Run KMeans with the remaining  $k$  centroids without repeating.
- **Step 4:** Decrease  $m$  by 1.
- **Step 5:** Repeat Steps 2 to 4 until  $m=0$ .

**Breathe in** step inserts new centroids close to the centroids with the largest errors. A centroid's error is the sum of the squared distance of points under that centroid.

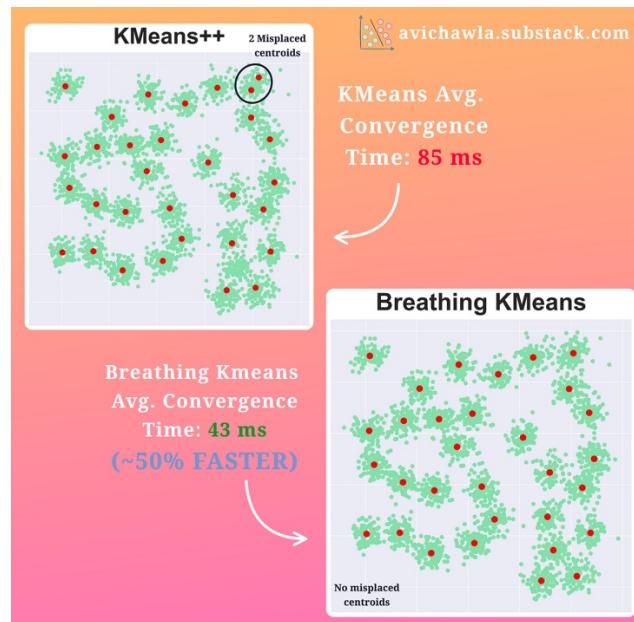
**Breathe out** step removes centroids with low utility. A centroid's utility is proportional to its distance from other centroids. The intuition is that if two centroids are pretty close, they are likely falling in the same cluster. Thus, both will be assigned a low utility value, as demonstrated below.



With these repeated breathing cycles, Breathing KMeans provides a faster and better solution than KMeans. In each cycle, new centroids are added at “good” locations, and centroids with low utility are removed.



In the figure below, KMeans++ produced two misplaced centroids.



However, Breathing KMeans accurately clustered the data, with a 50% improvement in run-time.

You can use Breathing KMeans by installing its open-source library, **bkmeans**, as follows:

```
pip install bkmeans
```

Next, import the library and run the clustering algorithm:

```
import numpy as np
from bkmeans import BKMeans

# generate random data set
X=np.random.rand(1000,2)

# create BKMeans instance
bkm = BKMeans(n_clusters=100)

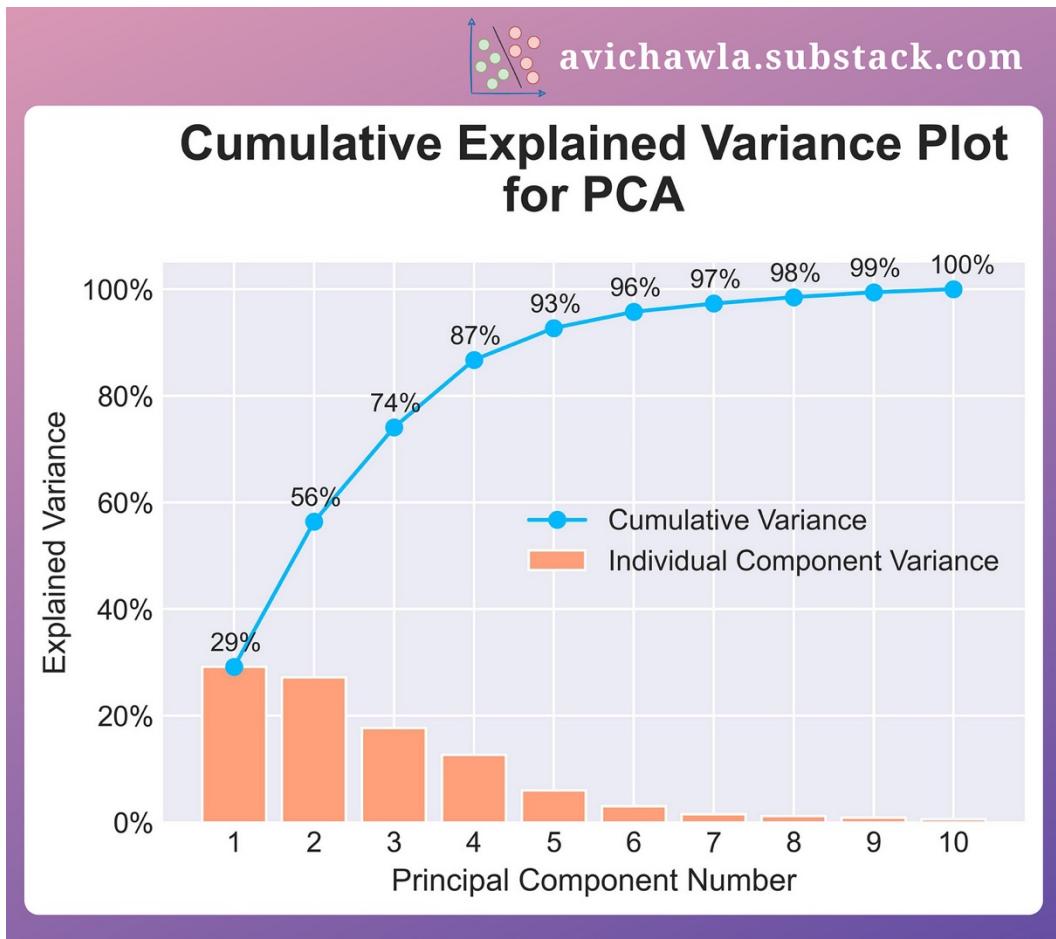
# run the algorithm
bkm.fit(X)
```

In fact, the `BKMeans` class inherits from the `KMeans` class of `sklearn`. So you can specify other parameters and use any of the other methods on the `BKMeans` object as needed.

More details about Breathing KMeans: [GitHub](#) | [Paper](#).



# How Many Dimensions Should You Reduce Your Data To When Using PCA?

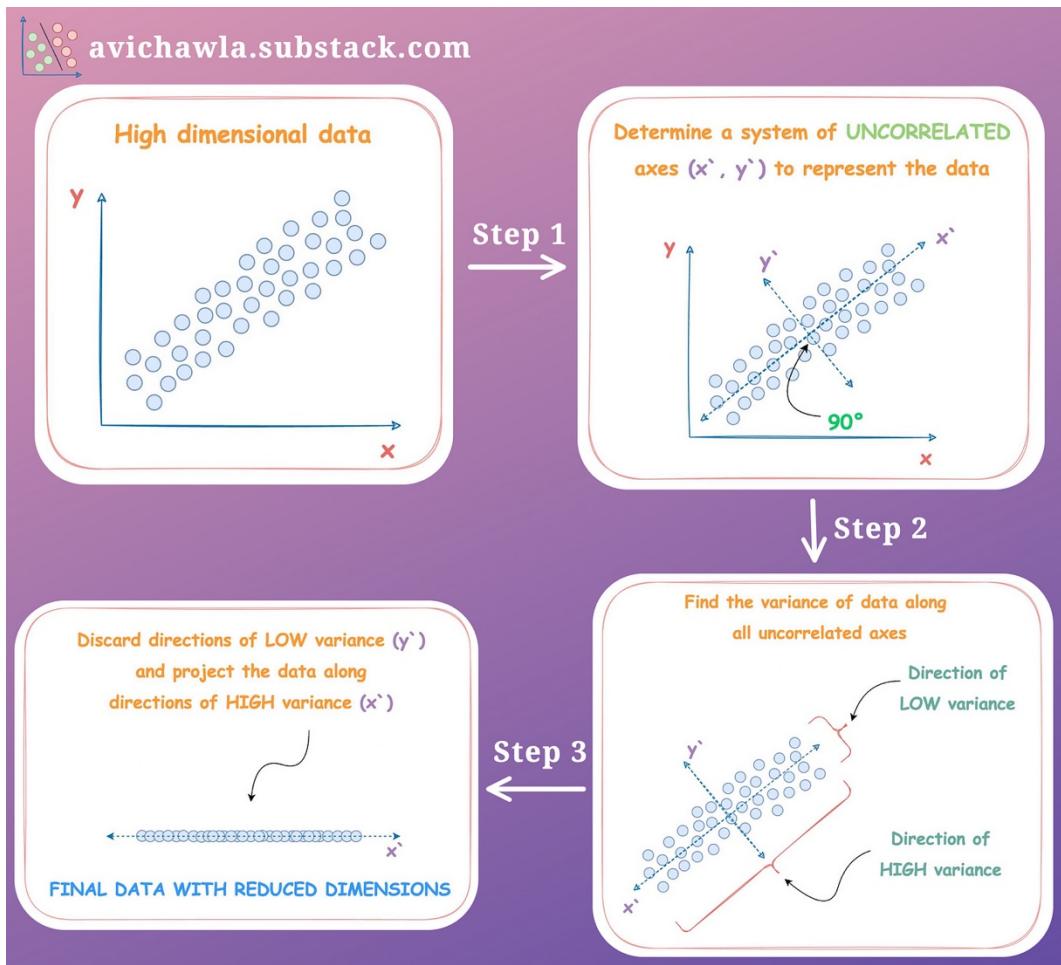


When using PCA, it can be difficult to determine the number of components to keep. Yet, here's a plot that can immensely help.

---

Note: If you don't know how PCA works, feel free to read my detailed post: [A Visual Guide to PCA](#).

Still, here's a quick step-by-step refresher. Feel free to skip this part if you remember my PCA post.



**Step 1.** Take a high-dimensional dataset ( $(\mathbf{x}, \mathbf{y})$  in the above figure) and represent it with uncorrelated axes ( $(\mathbf{x}', \mathbf{y}')$  in the above figure). Why uncorrelated?

This is to ensure that data has zero correlation along its dimensions and each new dimension represents its individual variance.

For instance, as data represented along  $(\mathbf{x}, \mathbf{y})$  is correlated, the variance along  $\mathbf{x}$  is influenced by the spread of data along  $\mathbf{y}$ .

Instead, if we represent data along  $(\mathbf{x}', \mathbf{y}')$ , the variance along  $\mathbf{x}'$  is not influenced by the spread of data along  $\mathbf{y}'$ .

The above space is determined using eigenvectors.

**Step 2.** Find the variance along all uncorrelated axes ( $\mathbf{x}', \mathbf{y}'$ ). The eigenvalue corresponding to each eigenvector denotes the variance.

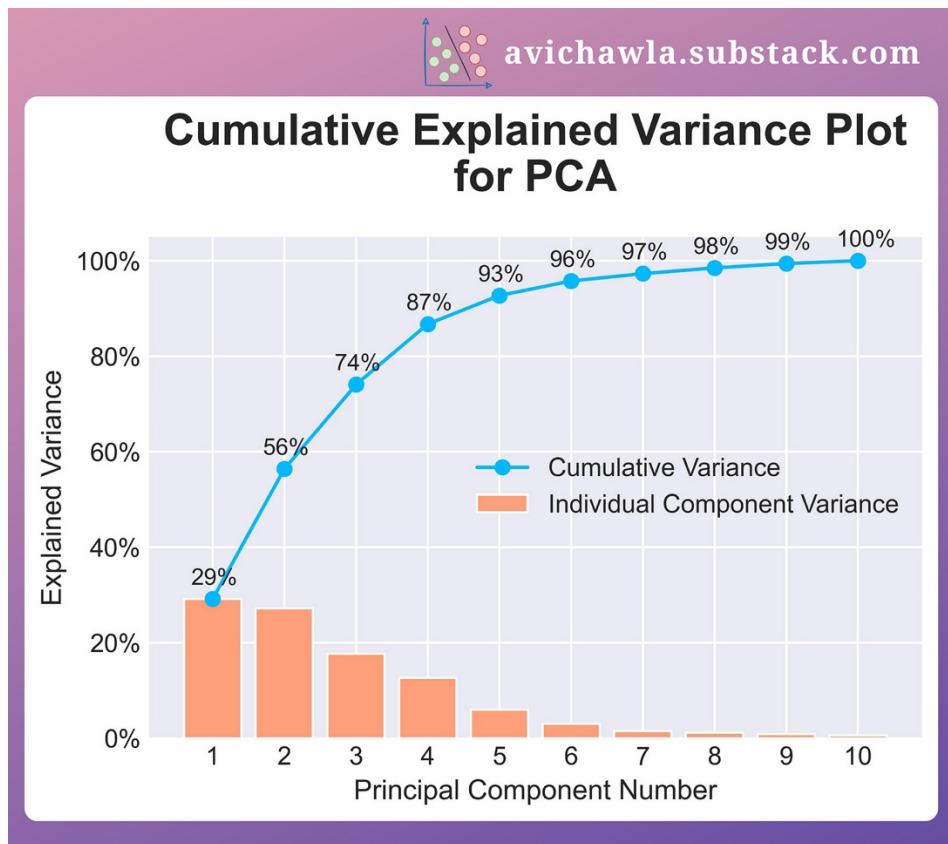
**Step 3.** Discard the axes with low variance. How many dimensions to discard (or keep) is a hyperparameter, which we will discuss below. Project the data along the retained axes.



When reducing dimensions, the purpose is to retain enough variance of the original data.

As each principal component explains some amount of variance, cumulatively plotting the component-wise variance can help identify which components have the most variance.

This is called a cumulative explained variance plot.



For instance, say we intend to retain **~85%** of the data variance. The above plot clearly depicts that reducing the data to four components will do that.

Also, as expected, all ten components together represent 100% variance of the data.

Creating this plot is pretty simple in Python. **Find the code here: [PCA-CEV Plot](#).**



# 🚀 Mito Just Got Supercharged With AI!

The screenshot shows the Mito spreadsheet interface. At the top, there's a toolbar with various icons for file operations like Undo, Redo, Import, Export, and a prominent AI icon. A green arrow points from the text above to this AI icon. Below the toolbar is a data table titled 'employee\_dataset' with 14 rows and 6 columns. The columns are labeled: Name, Company, City, Salary, Status, and Rating. The data includes names like Johnny Maynard, Michael Williams, and Linda Rodriguez, along with their respective company names and city locations. The salary column shows values like 8,804.92 and 12,505.38. The status column indicates full-time or intern positions, and the rating column shows scores ranging from 3.00 to 4.80. The bottom right corner of the table area says '(100 rows, 6 cols)'.

Personally, I am a big fan of no-code data analysis tools. They are extremely useful in eliminating repetitive code across projects—thereby boosting productivity.

Yet, most no-code tools are often limited in terms of the functionality they support. Thus, flexibility is usually a big challenge while using them.

Mito is an incredible open-source tool that allows you to analyze your data within a spreadsheet interface in Jupyter without writing any code.

What's more, Mito recently supercharged its spreadsheet interface with AI. As a result, you can now analyze data in a notebook with text prompts.

One of the coolest things about using Mito is that each edit in the spreadsheet automatically generates an equivalent Python code. This makes it convenient to reproduce the analysis later.



# Automatic code generation

```
from mitosheet.public.v3 import *; register_analysis("id-utbdzhhmhd");
import pandas as pd

# Imported employee_dataset.csv
employee_dataset = pd.read_csv(r'employee_dataset.csv')

# group on city and find avg salary and rating
df2 = employee_dataset.groupby('City').agg({'Salary': 'mean', 'Rating': 'mean'})

# top 5 employees with highest salary
top_employees = employee_dataset.nlargest(5, 'Salary')
```

You can install Mito using pip as follows:

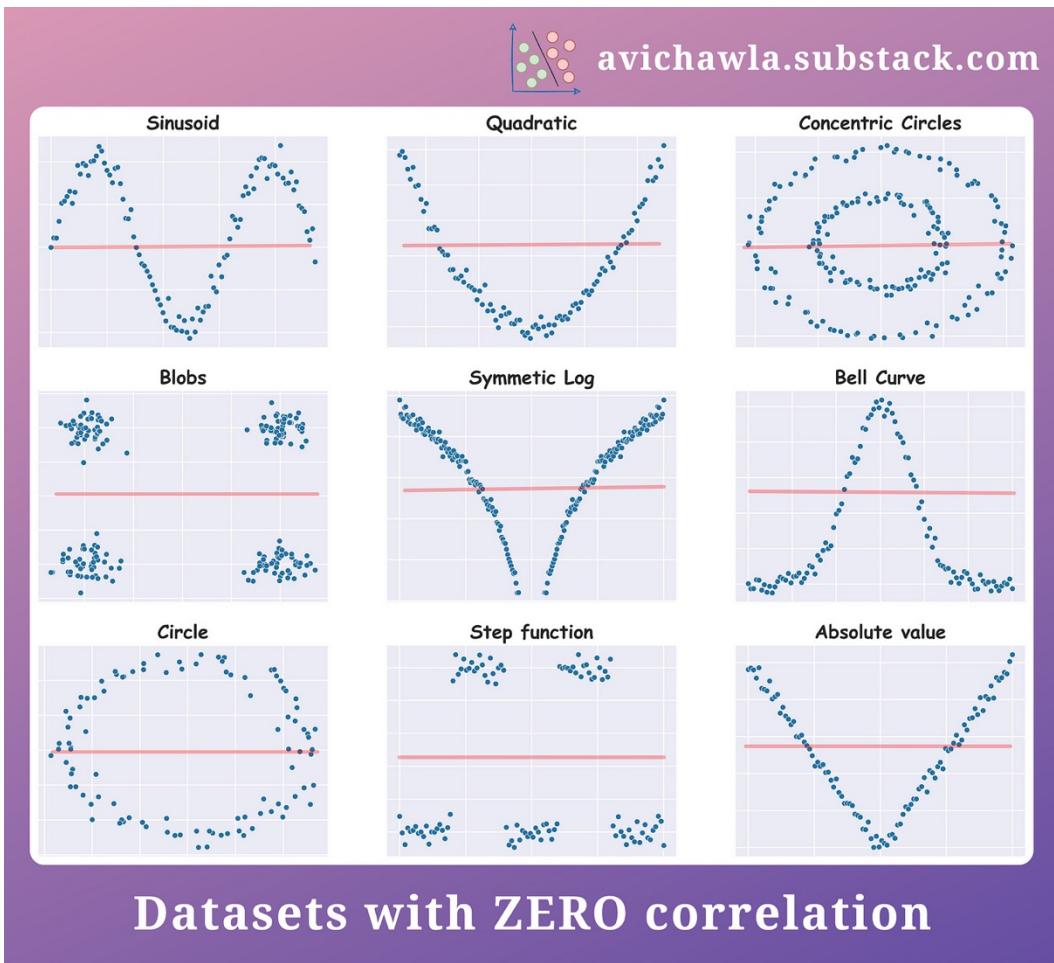
```
python -m pip install mitosheet
```

Next, to activate it in Jupyter, run the following two commands:

```
python -m jupyter nbextension install --py --user mitosheet
python -m jupyter nbextension enable --py --user mitosheet
```



# Be Cautious Before Drawing Any Conclusions Using Summary Statistics



Datasets with ZERO correlation

While analyzing data, one may be tempted to draw conclusions solely based on its statistics. Yet, the actual data might be conveying a totally different story.

Here's a visual depicting nine datasets with approx. zero correlation between the two variables. But the summary statistic (Pearson correlation in this case) gives no clue about what's inside the data.

What's more, data statistics could be heavily driven by outliers or other artifacts. I covered this in a previous post [here](#).

Thus, the importance of looking at the data cannot be stressed enough. It saves you from drawing wrong conclusions, which you could have made otherwise by looking at the statistics alone.

For instance, in the sinusoidal dataset above, Pearson correlation may make you believe that there is no association between the two variables. However, remember that it is only quantifying the extent of



[avichawla.substack.com](https://avichawla.substack.com)

a linear relationship between them. Read more about this in another one of my previous posts [here](#).

Thus, if there's any other non-linear relationship (quadratic, sinusoid, exponential, etc.), it will fail to measure that.



# Use Custom Python Objects In A Boolean Context

```
without_bool.py
```

```
class Cart:
    def __init__(self):
        self.items = []

# No __bool__ method

my_cart = Cart()

if my_cart:
    print("Cart Not Empty")
else:
    print("Cart Empty")

"Cart Not Empty" # Output
```

Object of custom class evaluated to True by default

```
with_bool.py
```

```
class Cart:
    def __init__(self):
        self.items = []

    def __bool__(self):
        return len(self.items)>0

my_cart = Cart()

if my_cart:
    print("Cart Not Empty")
else:
    print("Cart Empty")

"Cart Empty" # Output
```

Object evaluated to False

In a boolean context, Python always evaluates the objects of a custom class to True. But this may not be desired in all cases. Here's how you can override this behavior.

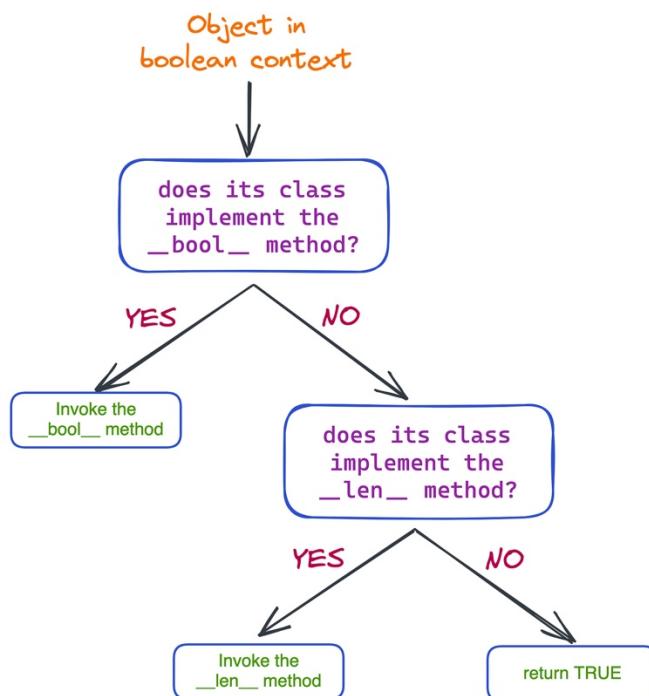
The `__bool__` dunder method is used to define the behavior of an object when used in a boolean context. As a result, you can specify explicit conditions to determine the truthiness of an object.

This allows you to use class objects in a more flexible and intuitive way.

As demonstrated above, without the `__bool__` method (`without_bool.py`), the object evaluates to True. But implementing the `__bool__` method lets us override this default behavior (`with_bool.py`).

## Some additional good-to-know details

When we use ANY object (be it instantiated from a custom or an in-built class) in a boolean context, here's what Python does:

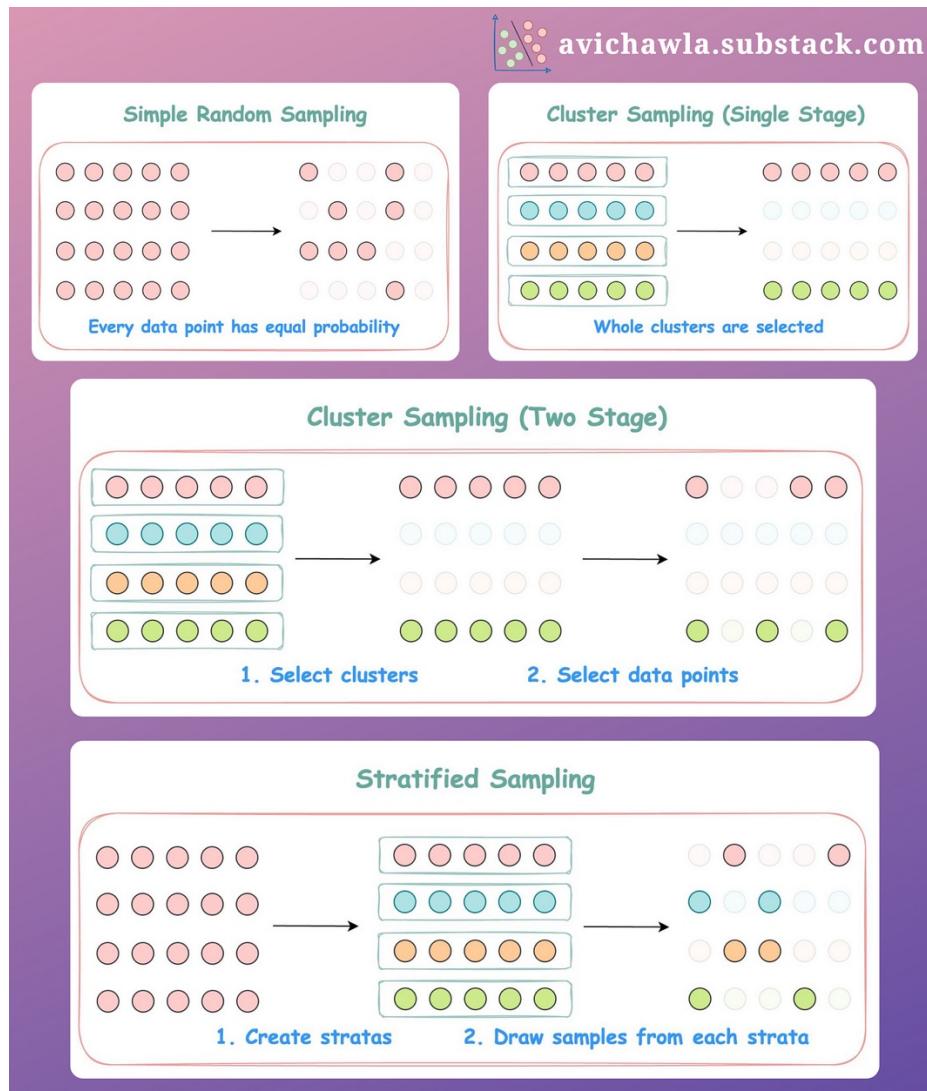


First, Python checks for the `__bool__` method in its class implementation. If found, it is invoked. If not, Python checks for the `__len__` method. If found, `__len__` is invoked. Otherwise, Python returns True.

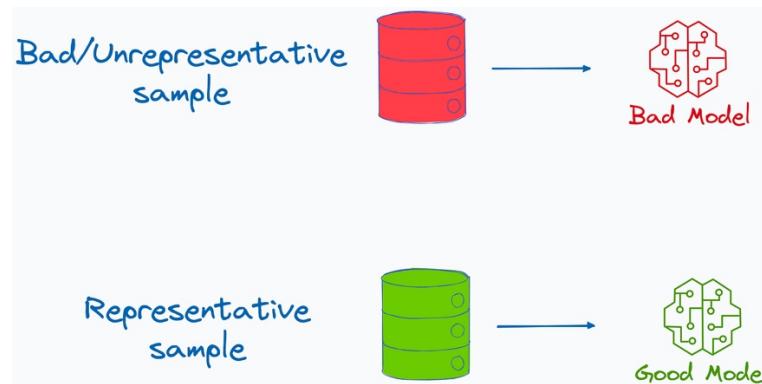
This explains the default behavior of objects instantiated from a custom class. As the `Cart` class implemented neither the `__bool__` method nor the `__len__` method, the `cart` object was evaluated to True.



# A Visual Guide To Sampling Techniques in Machine Learning



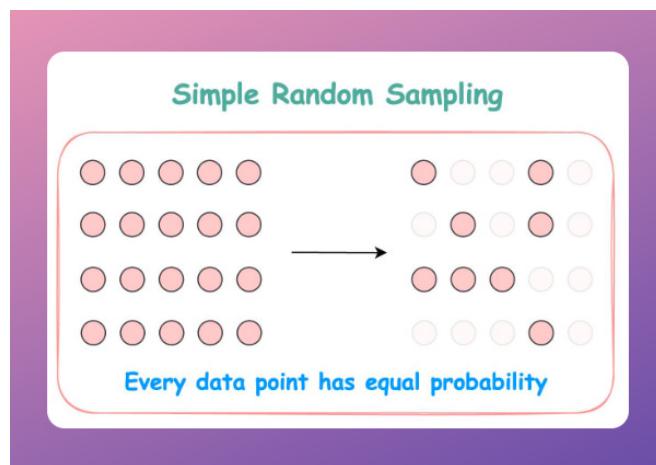
When you are dealing with large amounts of data, it is often preferred to draw a relatively smaller sample and train a model. But any mistakes can adversely affect the accuracy of your model.



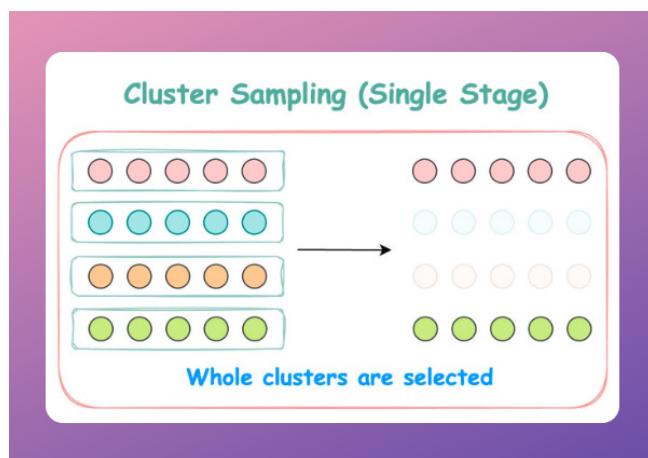
This makes sampling a critical aspect of training ML models.

Here are a few popularly used techniques that one should know about:

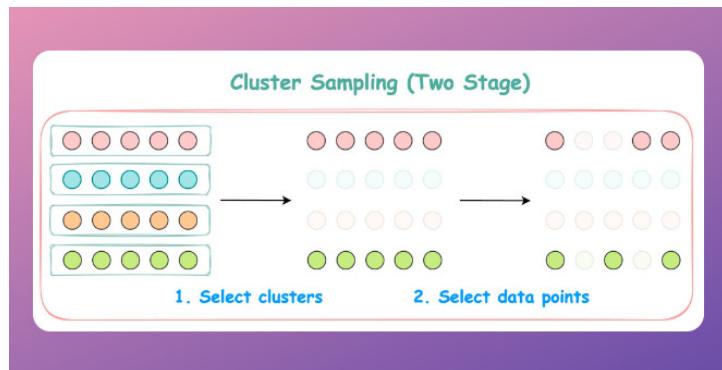
- ◆ **Simple random sampling:** Every data point has an equal probability of being selected in the sample.



- ◆ **Cluster sampling (single-stage):** Divide the data into clusters and select a few entire clusters.

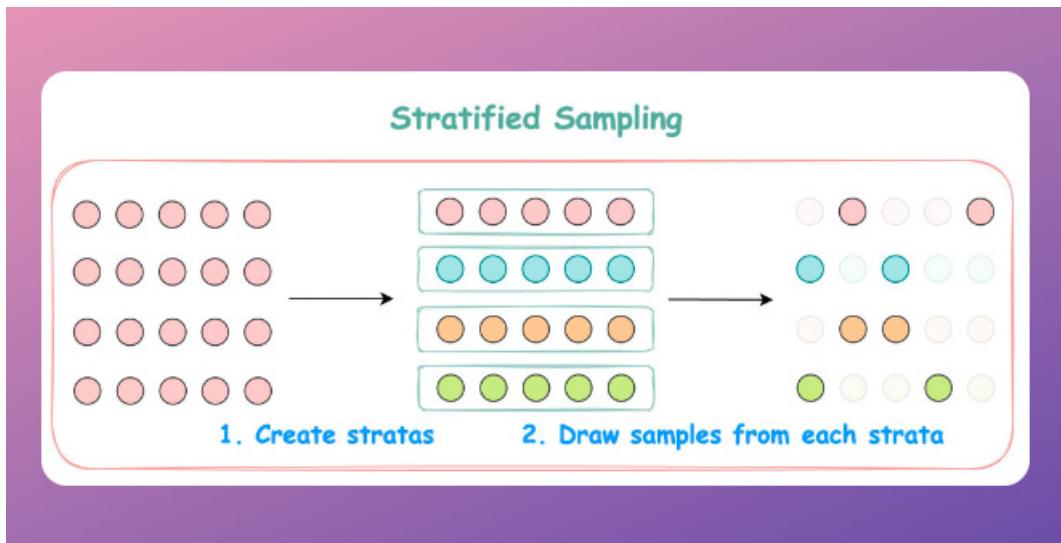


- ◆ **Cluster sampling (two-stage):** Divide the data into clusters, select a few clusters, and choose points from them randomly.





- ◆ **Stratified sampling:** Divide the data points into homogenous groups (based on age, gender, etc.), and select points randomly.



What are some other sampling techniques that you commonly resort to?



# You Were Probably Given Incomplete Info About A Tuple's Immutability

```
>>> my_tuple = (1, [2, 3])  
  
>>> my_tuple  
(1, [2, 3])  
  
>>> my_tuple[1].append(4) # No Error  
  
>>> my_tuple  
(1, [2, 3, 4])
```

Tuple Modified

When we say tuples are immutable, many Python programmers think that the values inside a tuple cannot change. But this is not true.

The immutability of a tuple is solely restricted to the identity of objects it holds, not their value.

In other words, say a tuple has two objects with IDs **1** and **2**. Immutability says that the collection of IDs referenced by the tuple (and their order) can never change.

Yet, there is **NO** such restriction that the individual objects with IDs **1** and **2** cannot be modified.

Thus, if the elements inside the tuple are mutable objects, you can indeed modify them.

And as long as the collection of IDs remains the same, the immutability of a tuple is not violated.



This explains the demonstration above. As `append` is an inplace operation, the collection of IDs didn't change. Thus, Python didn't raise an error.

We can also verify this by printing the collection of object IDs referenced inside the tuple before and after the append operation:

```
>>> my_tuple = (1, [2, 3])  
>>> id(my_tuple[0]), id(my_tuple[1])  
(583145, 434810)  
  
>>> my_tuple[1].append(4)  
>>> id(my_tuple[0]), id(my_tuple[1])  
(583145, 434810)
```

As shown above, the IDs pre and post append are the same. Thus, immutability isn't violated.



# A Simple Trick That Significantly Improves The Quality of Matplotlib Plots



Matplotlib plots often appear dull and blurry, especially when scaled or zoomed. Yet, here's a simple trick to significantly improve their quality.

Matplotlib plots are rendered as an image by default. Thus, any scaling/zooming drastically distorts their quality.

Instead, always render your plot as a scalable vector graphic (SVG). As the name suggests, they can be scaled without compromising the plot's quality.

As demonstrated in the image above, the plot rendered as SVG clearly outshines and is noticeably sharper than the default plot.



The following code lets you change the render format to SVG. If the difference is not apparent in the image above, I would recommend trying it yourself and noticing the difference.

```
from matplotlib_inline.backend_inline import set_matplotlib_formats  
set_matplotlib_formats('svg')
```

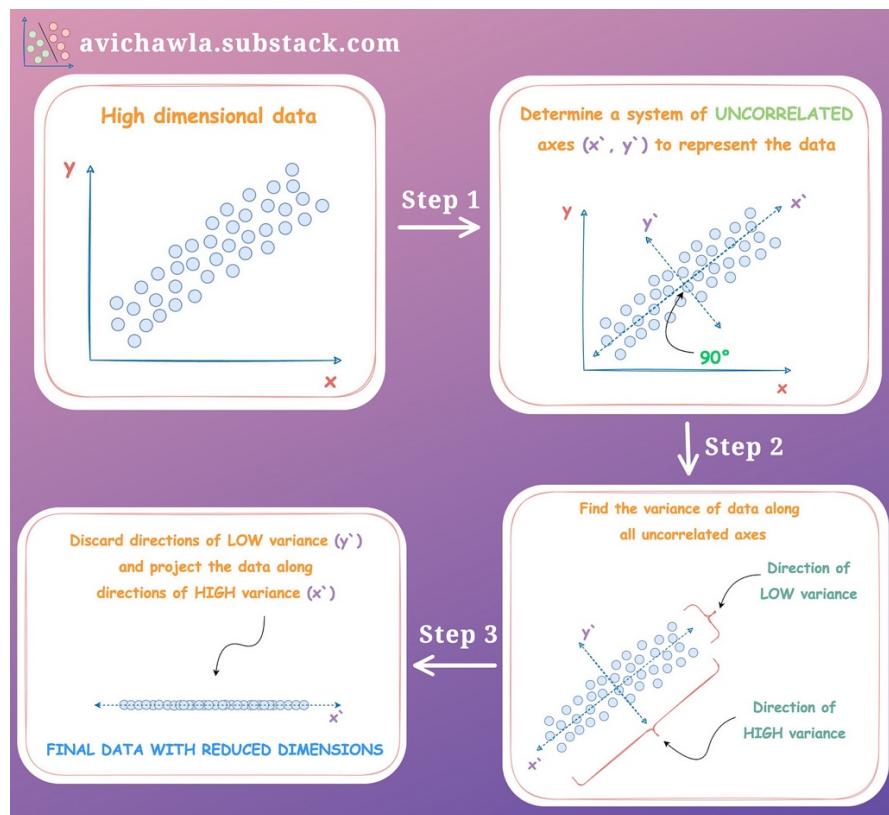
Alternatively, you can also use the following code:

```
%config InlineBackend.figure_format = 'svg'
```

P.S. If there's a chance that you don't know what is being depicted in the bar plot above, check out this [YouTube video by Numberphile](#).



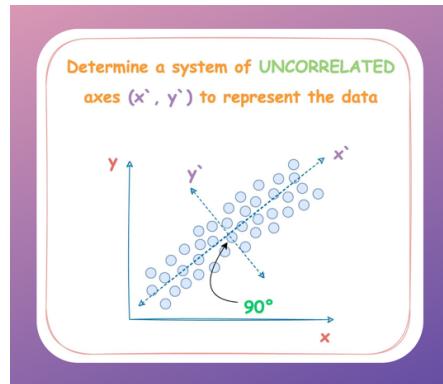
# A Visual and Overly Simplified Guide to PCA



Many folks often struggle to understand the core essence of principal component analysis (PCA), which is widely used for dimensionality reduction. Here's a simplified visual guide depicting what goes under the hood.

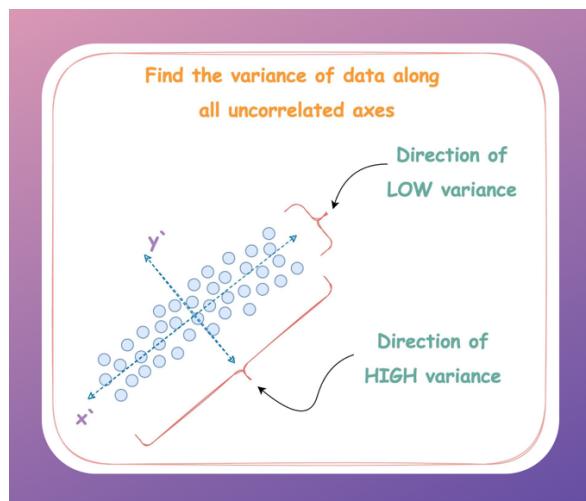
In a gist, while reducing the dimensions, the aim is to retain as much variation in data as possible.

To begin with, as the data may have correlated features, the first step is to determine a new coordinate system with orthogonal axes. This is a space where all dimensions are uncorrelated.



The above space is determined using the data's eigenvectors.

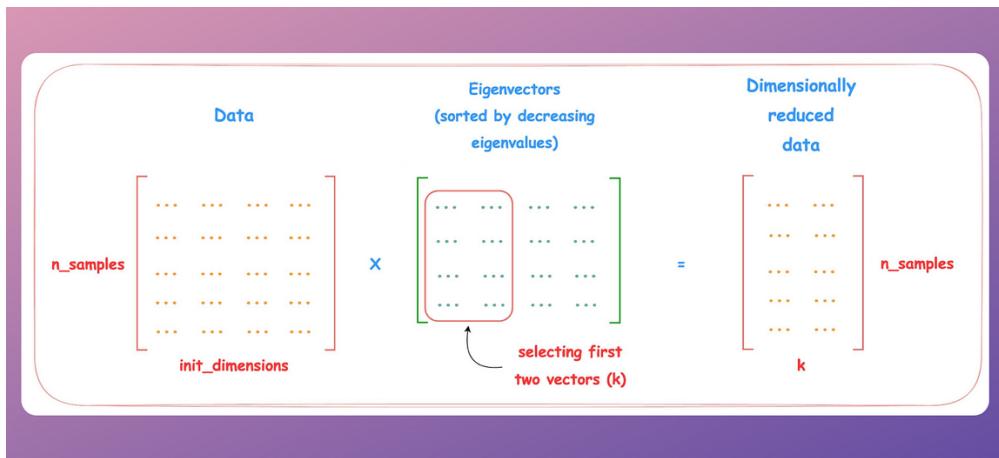
Next, we find the variance of our data along these uncorrelated axes. The variance is represented by the corresponding eigenvalues.



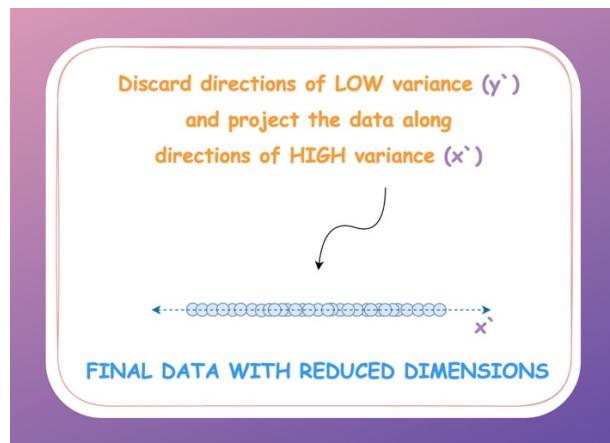
Next, we decide the number of dimensions we want our data to have post-reduction (a hyperparameter), say two. As our aim is to retain as much variance as possible, we select two eigenvectors with the highest eigenvalues.

Why highest, you may ask? As mentioned above, the variance along an eigenvector is represented by its eigenvalue. Thus, selecting the top two eigenvalues ensures we retain the maximum variance of the overall data.

Lastly, the data is transformed using a simple matrix multiplication with the top two vectors, as shown below:



After reducing the dimension of the 2D dataset used above, we get the following.



This is how PCA works. I hope this algorithm will never feel daunting again :)



# Supercharge Your Jupyter Kernel With ipyflow

This is a pretty cool Jupyter hack I learned recently.

While using Jupyter, you must have noticed that when you update a variable, all its dependent cells have to be manually re-executed.

Also, at times, isn't it difficult to determine the exact sequence of cell executions that generated an output?

This is tedious and can get time-consuming if the sequence of dependent cells is long.

To resolve this, try **ipyflow**. It is a supercharged kernel for Jupyter, which tracks the relationship between cells and variables.

```
In [1]: import numpy as np

Automatic Execution of Dependent Cells

In [2]: %flow mode reactive

In [ ]: x = 10 ## Updating x automatically executes its dependents

In [ ]: y = np.sin(x) ## Dependent on x
z = np.cos(x) ## Dependent on x

In [ ]: output = y**2 + z**2 ## Dependent on y and z
output

Export Code

In [ ]: from ipyflow import code
print(code(output))

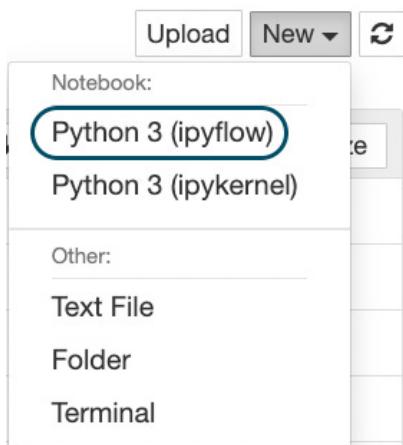
In [ ]:
```

Thus, at any point, you can obtain the corresponding code to reconstruct any symbol.

What's more, its magic command enables an automatic recursive re-execution of dependent cells if a variable is updated.

As shown in the demo above, updating the variable X automatically triggers its dependent cells.

Do note that **ipyflow** offers a different kernel from the default kernel in Jupyter. Thus, once you install **ipyflow**, select the following kernel while launching a new notebook:



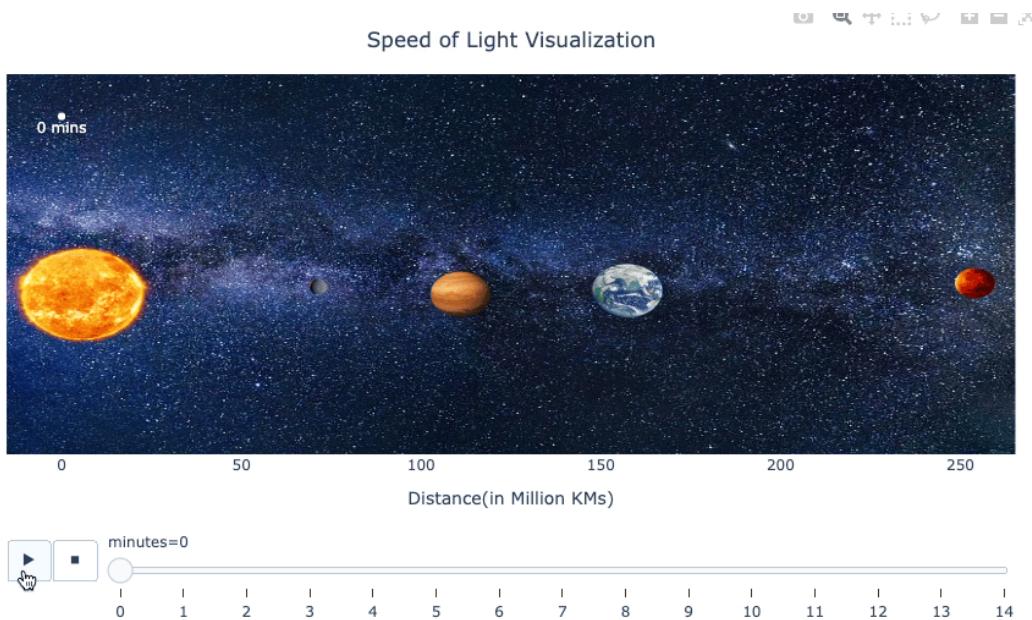
Find more details here: [ipyflow](#).



# A Lesser-known Feature of Creating Plots with Plotly

Plotly is pretty diverse when it comes to creating different types of charts. While many folks prefer it for interactivity, you can also use it to create animated plots.

Here's an animated visualization depicting the time taken by light to reach different planets after leaving the Sun.



Several functions in Plotly support animations using the **animation\_frame** and **animation\_group** parameters.

The core idea behind creating an animated plot relies on plotting the data one frame at a time.

For instance, consider we have organized the data frame-by-frame, as shown below:



	planets	x_position	y_position	frame_id
0	Sun	0.0	0.0	0
1	Mercury	70.0	0.0	0
2	Venus	110.0	0.0	0
3	Earth	150.0	0.0	0
4	Mars	250.0	0.0	0
5	Light	0.0	0.2	0

	planets	x_position	y_position	frame_id
6	Sun	0.0	0.0	1
7	Mercury	70.0	0.0	1
8	Venus	110.0	0.0	1
9	Earth	150.0	0.0	1
10	Mars	250.0	0.0	1
11	Light	18.0	0.2	1

...

Location in each frame

Now, if we invoke the scatter method with the **animation\_frame** argument, it will plot the data frame-by-frame, giving rise to an animation.

```
import plotly.express as px

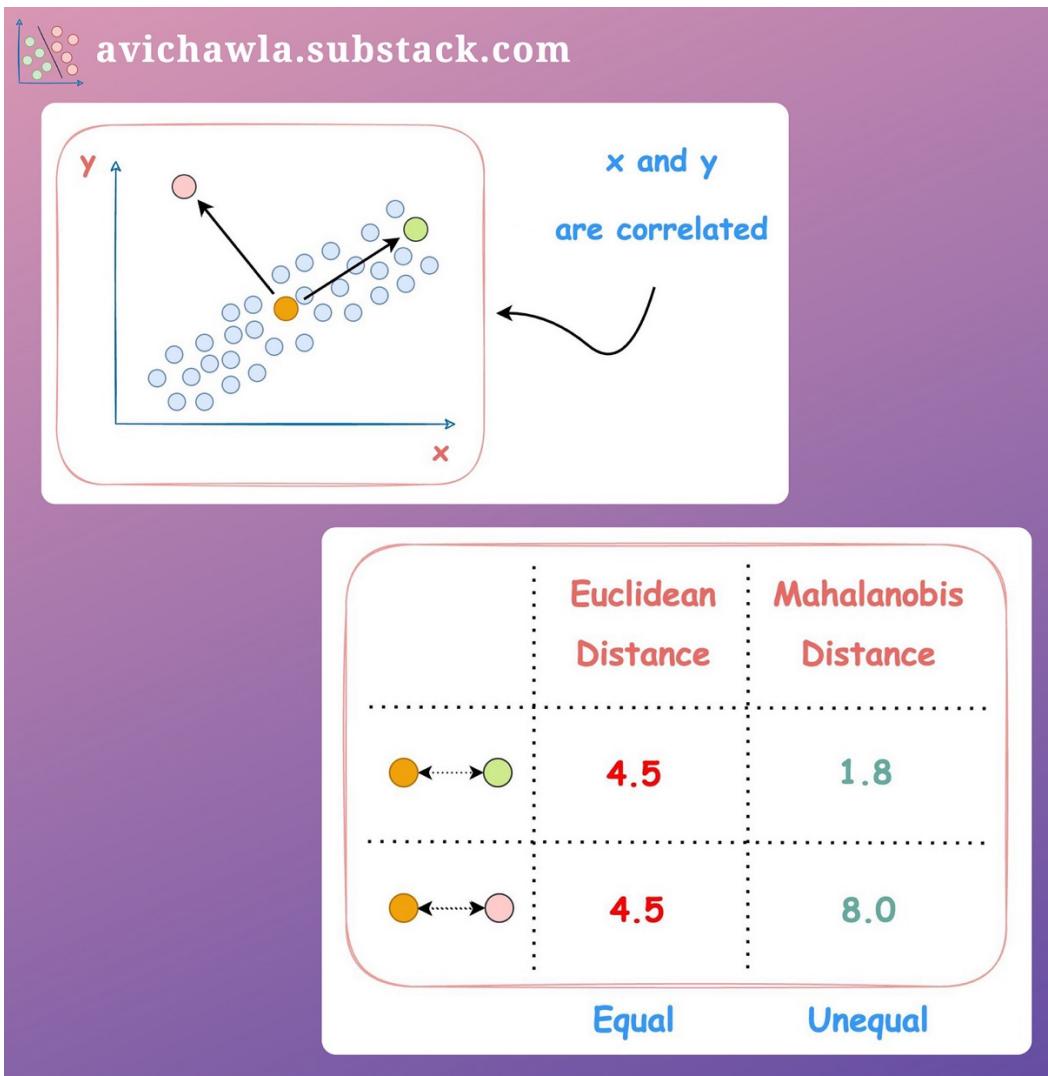
>>> px.scatter(df,
      x="x_position",
      y="y_position",
      color = "planets",
      animation_frame="frame_id")
```

In the above function call, the data corresponding to **frame\_id=0** will be plotted first. This will be replaced by the data with **frame\_id=1** in the next frame, and so on.

Find the code for this post here: [GitHub](#).



# The Limitation Of Euclidean Distance Which Many Often Ignore



Euclidean distance is a commonly used distance metric. Yet, its limitations often make it inapplicable in many data situations.

Euclidean distance assumes independent axes, and the data is somewhat spherically distributed. But when the dimensions are correlated, euclidean may produce misleading results.

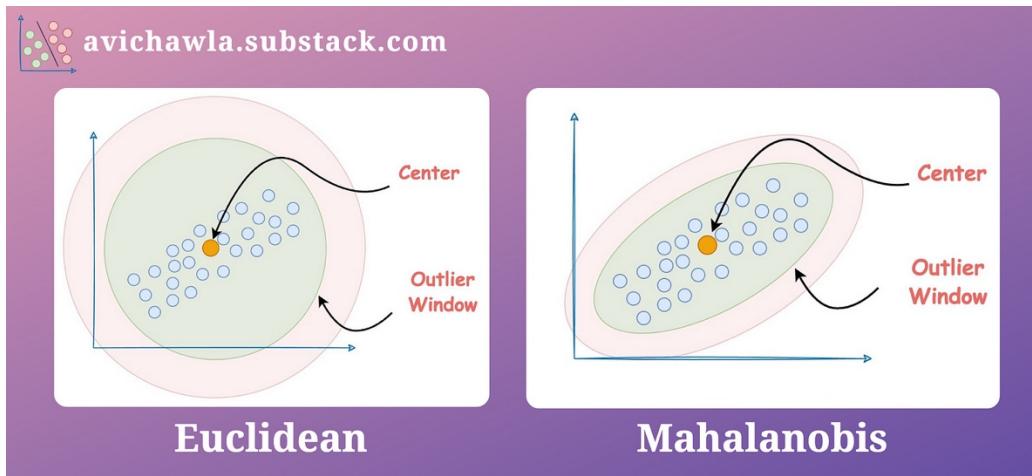
Mahalanobis distance is an excellent alternative in such cases. It is a multivariate distance metric that takes into account the data distribution.

As a result, it can measure how far away a data point is from the distribution, which Euclidean cannot.



As shown in the image above, Euclidean considers pink and green points equidistant from the central point. But Mahalanobis distance considers the green point to be closer, which is indeed true, taking into account the data distribution.

Mahalanobis distance is commonly used in outlier detection tasks. As shown below, while Euclidean forms a circular boundary for outliers, Mahalanobis, instead, considers the distribution—producing a more practical boundary.



Essentially, Mahalanobis distance allows the data to construct a coordinate system for itself, in which the axes are independent and orthogonal.

Computationally, it works as follows:

- **Step 1:** Transform the columns into uncorrelated variables.
- **Step 2:** Scale the new variables to make their variance equal to 1.
- **Step 3:** Find the Euclidean distance in this new coordinate system, where the data has a unit variance.

So eventually, we do reach Euclidean. However, to use Euclidean, we first transform the data to ensure it obeys the assumptions.

Mathematically, it is calculated as follows:

$$D^2 = (x - \mu)^T \cdot C^{-1} \cdot (x - \mu)$$

- $x$ : rows of your dataset (Shape:  $n\_samples \times n\_dimensions$ ).
- $\mu$ : mean of individual dimensions (Shape:  $1 \times n\_dimensions$ ).



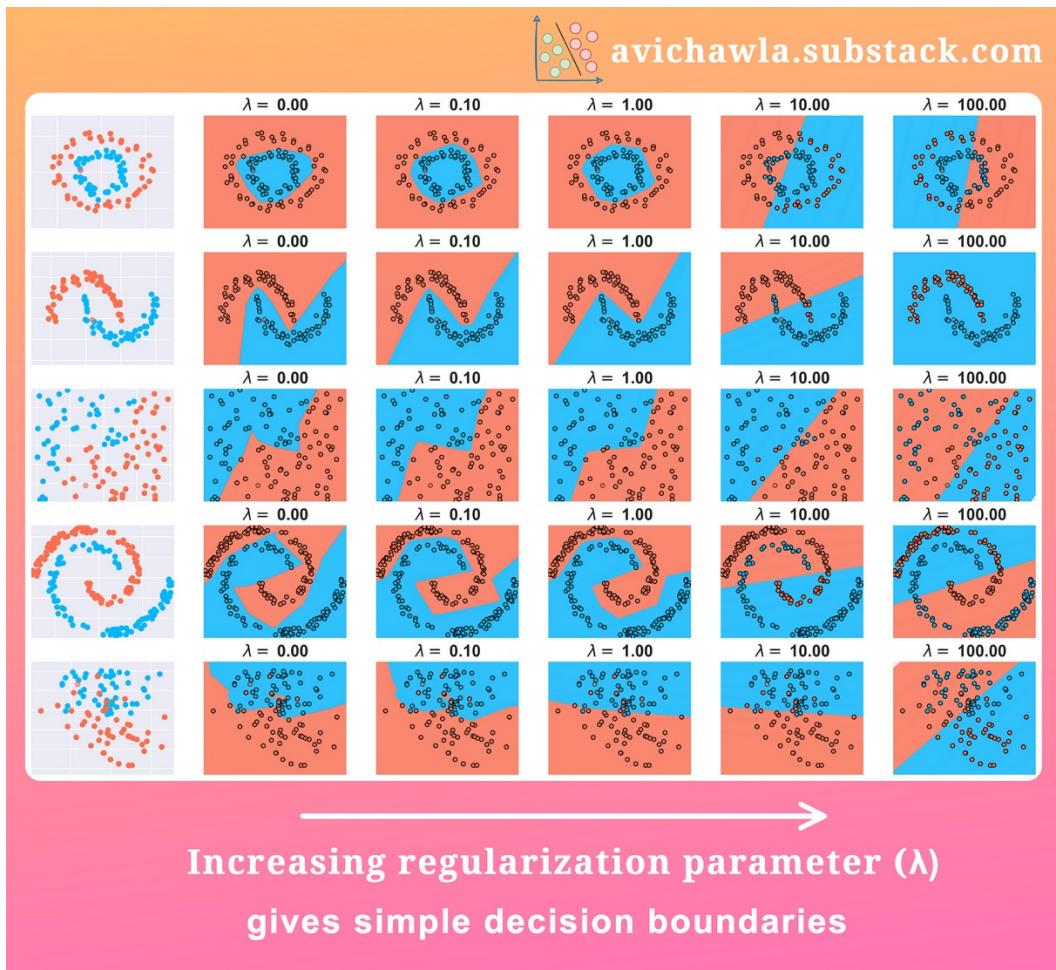
[avichawla.substack.com](https://avichawla.substack.com)

- $C^{-1}$ : Inverse of the covariance matrix  
(Shape: n\_dimensions\*n\_dimensions).
- $D^2$ : Square of the Mahalanobis distance  
(Shape: n\_samples\*n\_samples).

Find more info here: [Scipy docs](#).



# Visualising The Impact Of Regularisation Parameter



Regularization is commonly used to prevent overfitting. The above visual depicts the decision boundary obtained on various datasets by varying the regularization parameter.

As shown, increasing the parameter results in a decision boundary with fewer curvatures. Similarly, decreasing the parameter produces a more complicated decision boundary.

But have you ever wondered what goes on behind the scenes? Why does increasing the parameter force simpler decision boundaries?

To understand that, consider the cost function equation below (this is for regression though, but the idea stays the same for classification).

It is clear that the cost increases linearly with the parameter  $\lambda$ .



Cost Function = Loss + L2 Weight Penalty

$$= \underbrace{\sum_{i=1}^M (y_i - \sum_{j=1}^N x_{ij} w_j)^2}_{\text{Squared Error}} + \lambda \underbrace{\sum_{j=1}^N w_j^2}_{\text{L2 Regularization Term}}$$

**Higher the value of  $\lambda$ ,  
higher the penalty**

Now, if the parameter is too high, the penalty becomes higher too. Thus, to minimize its impact on the overall cost function, the network is forced to approach weights that are closer to zero.

This becomes evident if we print the final weights for one of the models, say one at the bottom right (last dataset, last model).

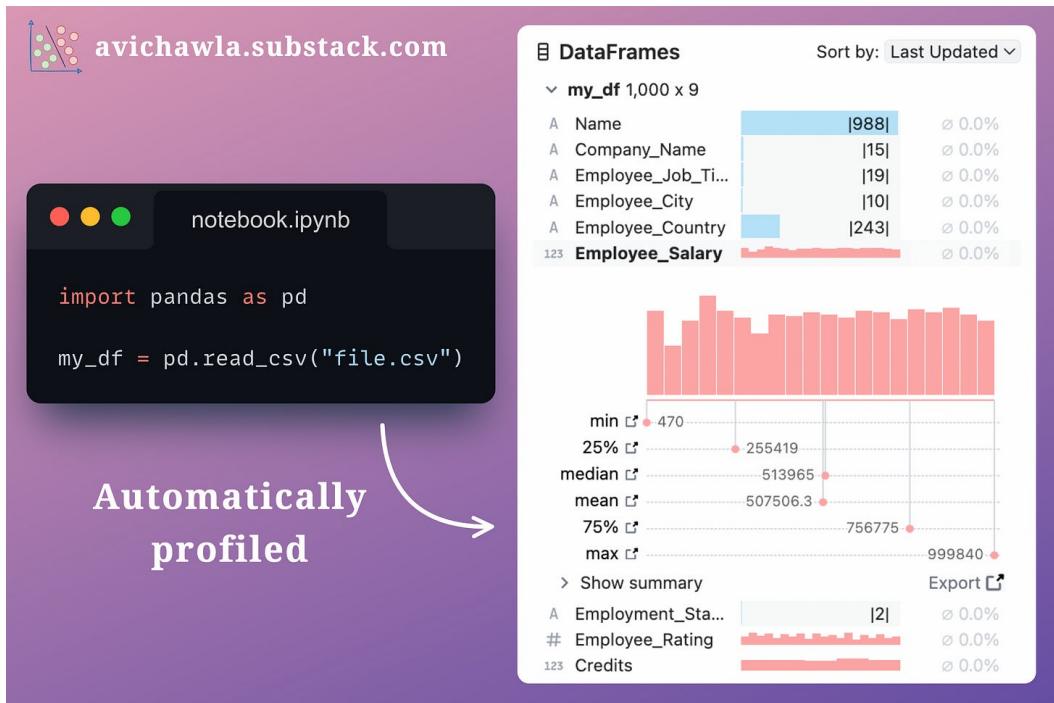
**All weights close to zero**

```
In [17]: clf.coefs_
Out[17]: array([[ 8.35476806e-06, -1.29066987e-05,  1.49535843e-05,
   8.43964067e-06,  5.46943218e-06,  1.18557175e-05,
  1.01037005e-05,  3.70503012e-06,  2.12142850e-06,
 -9.78452613e-06],
 [-1.35980250e-05,  1.52132934e-05,  3.30938991e-06,
  7.41538247e-07,  1.68626879e-05,  1.14315983e-05,
  6.64292409e-07, -1.40798113e-06,  1.31551207e-05,
  2.52379486e-05]])
```

Having smaller weights effectively nullifies many neurons, producing a much simpler network. This prevents many complex transformations, that could have happened otherwise.



# AutoProfiler: Automatically Profile Your DataFrame As You Work



Pandas AutoProfiler: Automatically profile Pandas DataFrames at each execution, without any code.

AutoProfiler is an open-source dataframe analysis tool in jupyter. It reads your notebook and automatically profiles every dataframe in your memory as you change them.

In other words, if you modify an existing dataframe, AutoProfiler will automatically update its corresponding profiling.

Also, if you create a new dataframe (say from an existing dataframe), AutoProfiler will automatically profile that as well, as shown below:



New DataFrame

Profile

```
import pandas as pd  
my_df = pd.read_csv("file.csv")  
  
new_df = my_df.sample(100)
```

DataFrames

Column	Type	Count	Null %
Name	A	100	0.0%
Company_Name	A	15	0.0%
Employee_Job_Ti...	A	19	0.0%
Employee_City	A	10	0.0%
Employee_Country	A	85	0.0%
Employee_Salary	123	123	0.0%
Employment_Sta...	A	2	0.0%
Employee_Rating	#	123	0.0%
Credits	123	123	0.0%

Sort by: Last Updated ▾

Profiling info includes column distribution, summary stats, null stats, and many more. Moreover, you can also generate the corresponding code, with its export feature.

Code added in cell

Export code

```
import pandas as pd  
my_df = pd.read_csv("file.csv")  
  
my_df[my_df["Name"] == "Sarah Smith"]
```

DataFrames

Column	Type	Count	Null %
Name	A	988	0.0%
Kelly Young	2	(0.20%)	
Renee Davis	2	(0.20%)	
Patty Jones	2	(0.20%)	
William Jones	2	(0.20%)	
Daniel Lee	2	(0.20%)	
Sarah Smith	2	(0.20%)	
Anna Thomas	2	(0.20%)	
Jennifer Gonzales	2	(0.20%)	
Nicole Garcia	2	(0.20%)	
Derek Perez	2	(0.20%)	

Show summary

Export ↗

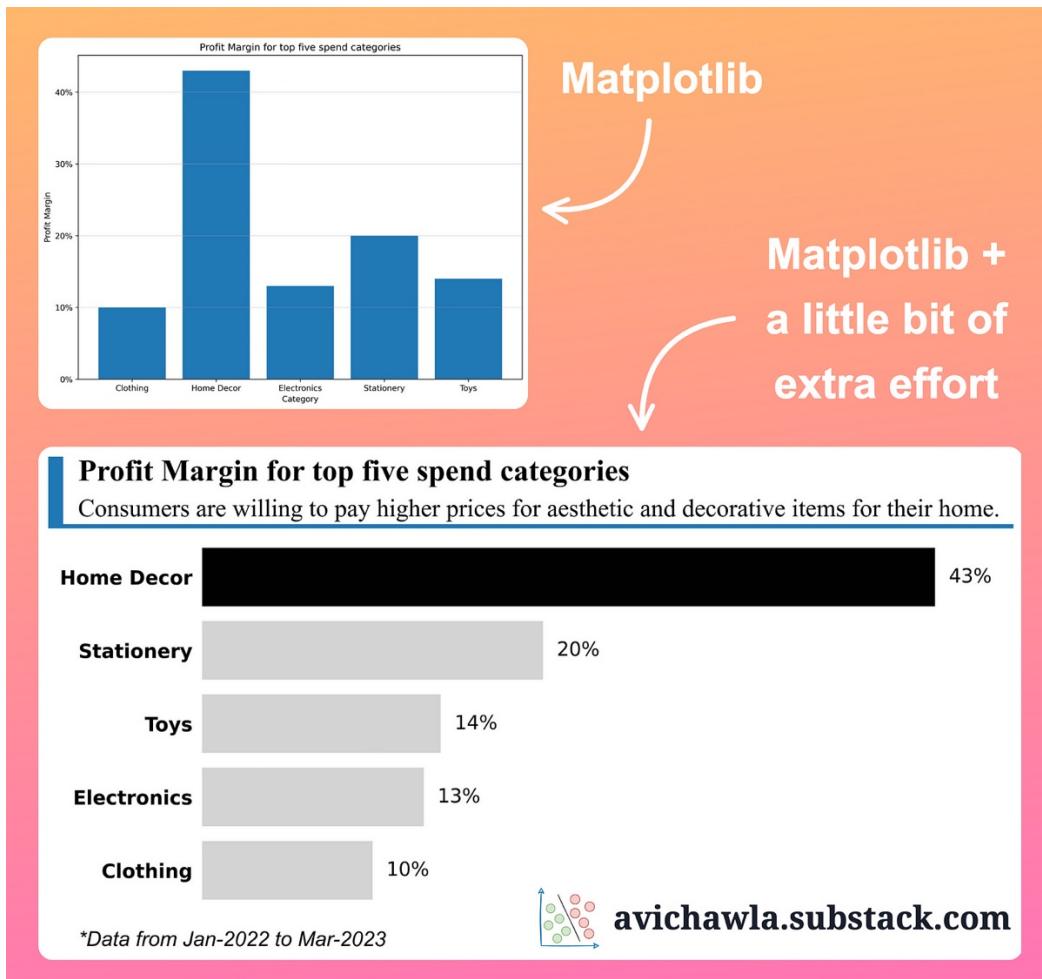
Column	Type	Count	Null %
Company_Name	A	15	0.0%
Employee_Job_Ti...	A	19	0.0%
Employee_City	A	10	0.0%
Employee_Country	A	243	0.0%
Employee_Salary	123	123	0.0%
Employment_Sta...	A	2	0.0%
Employee_Rating	#	123	0.0%
Credits	123	123	0.0%

Sort by: Last Updated ▾

Find more info here: [GitHub Repo.](#)



# A Little Bit Of Extra Effort Can Hugely Transform Your Storytelling Skills



Matplotlib is pretty underrated when it comes to creating professional-looking plots. Yet, it is totally capable of doing so.

For instance, consider the two plots below.

Yes, both were created using matplotlib. But a bit of formatting makes the second plot much more informative, appealing, and easy to follow.

The title and subtitle significantly aid the story. Also, the footnote offers extra important information, which is nowhere to be seen in the basic plot.

Lastly, the bold bar immediately draws the viewer's attention and conveys the category's importance.

So what's the message here?



[avichawla.substack.com](https://avichawla.substack.com)

Towards being a good data storyteller, ensure that your plot demands minimal effort from the viewer. Thus, don't shy away from putting in that extra effort. This is especially true for professional environments.

At times, it may be also good to ensure that your visualizations convey the right story, even if they are viewed in your absence.



# A Nasty Hidden Feature of Python That Many Programmers Aren't Aware Of

**Mutable Default Parameter**

```
def add_subject(name, subject, subjects=[]):
    subjects.append(subject)
    return {'name': name, 'subjects': subjects}

>>> add_subject('Joe', 'Maths')
>>> add_subject('Bob', 'Maths')
>>> add_subject('Roy', 'Maths')
```

**Output:**

```
{'name': 'Joe', 'subjects': ['Maths']}
{'name': 'Bob', 'subjects': ['Maths', 'Maths']}
{'name': 'Roy', 'subjects': ['Maths', 'Maths', 'Maths']}
```

**Appended to the same list**

avichawla.substack.com

Mutability in Python is possibly one of the most misunderstood and overlooked concepts. The above image demonstrates an example that many Python programmers (especially new ones) struggle to understand.

Can you figure it out? If not, let's understand it.

The default parameters of a function are evaluated right at the time the function is defined. In other words, they are not evaluated each time the function is called (like in C++).

Thus, as soon as a function is defined, the **function object** stores the default parameters in its `__defaults__` attribute. We can verify this below:



```
def my_function(a=1, b=2, c=3):
    pass

>>> my_function.__defaults__
(1, 2, 3)
```

Thus, if you specify a mutable default parameter in a function and mutate it, you unknowingly and unintentionally modify the parameter for all future calls to that function.

This is shown in the demonstration below. Instead of creating a new list at each function call, Python appends the element to the same copy.

Modified  
default  
parameter

```
def add_subject(...):
    ...

>>> add_subject.__defaults__
([],)

>>> add_subject('Joe', 'Maths')
>>> add_subject.__defaults__
(['Maths'],)

>>> add_subject('Bob', 'Maths')
>>> add_subject.__defaults__
(['Maths', 'Maths'])

>>> add_subject('Roy', 'Maths')
>>> add_subject.__defaults__
(['Maths', 'Maths', 'Maths'])
```



## So what can we do to avoid this?

Instead of specifying a mutable default parameter in a function's definition, replace them with None. If the function does not receive a corresponding value during the function call, create the mutable object inside the function.

This is demonstrated below:

**Replace mutable parameter**

```
def add_subject(name, subject, subjects=None):
    if subjects is None:
        # Create if no value was received
        subjects = []

    subjects.append(subject)
    return {'name': name, 'subjects': subjects}

>>> add_subject('Joe', 'Maths')
>>> add_subject('Bob', 'Maths')
>>> add_subject('Roy', 'Maths')
```

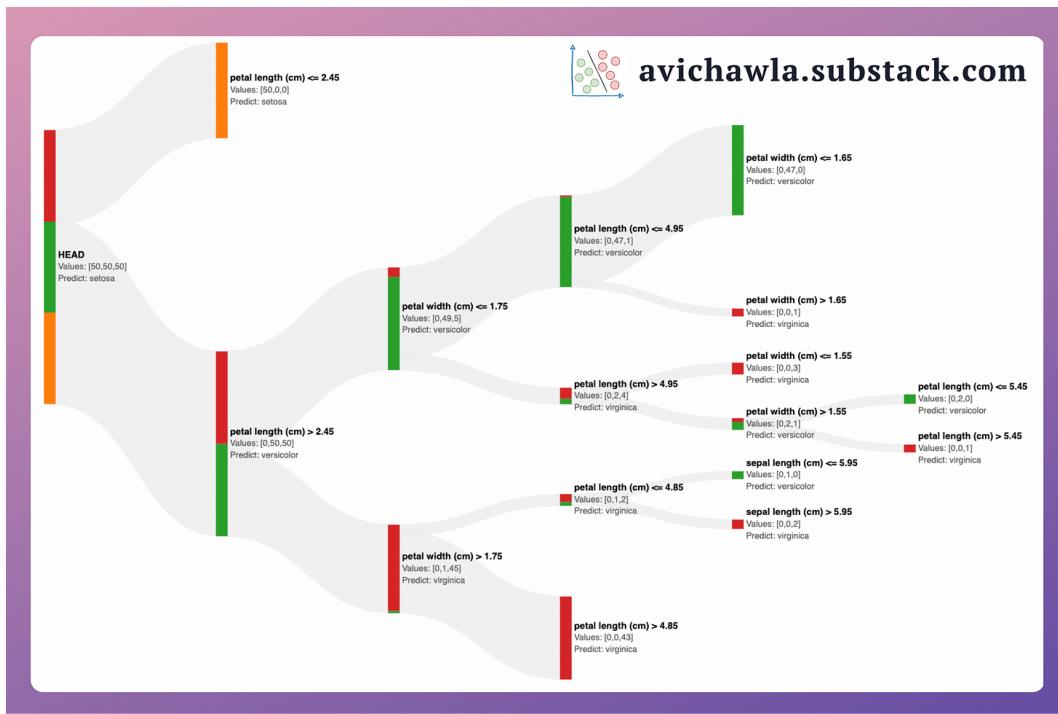
Output:

```
{'name': 'Joe', 'subjects': ['Maths']}
{'name': 'Bob', 'subjects': ['Maths']}
{'name': 'Roy', 'subjects': ['Maths']}
```

As shown above, we create a new list if the function didn't receive any value when it was called. This lets you avoid the unexpected behavior of mutating the same object.



# Interactively Visualise A Decision Tree With A Sankey Diagram



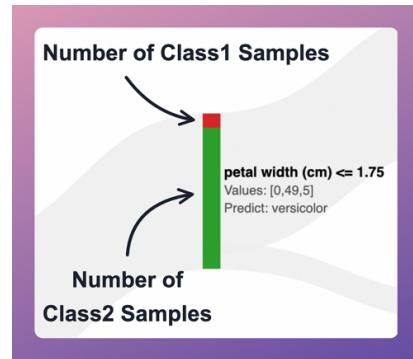
In one of my earlier posts, I explained why sklearn's decision trees always overfit the data with its default parameters (read [here](#) if you wish to recall).

To avoid this, it is always recommended to specify appropriate hyperparameter values. This includes the max depth of the tree, min samples in leaf nodes, etc.

But determining these hyperparameter values is often done using trial-and-error, which can be a bit tedious and time-consuming.

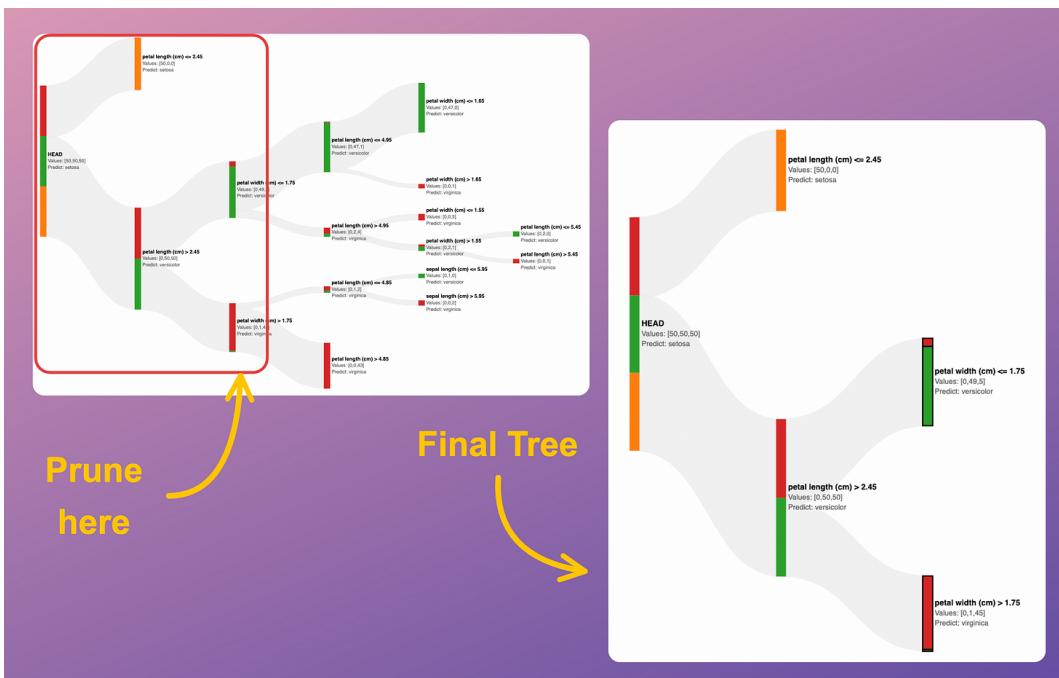
The Sankey diagram above allows you to interactively visualize the predictions of a decision tree at each node.

Also, the number of data points from each class is size-encoded on all nodes, as shown below.



This immediately gives an estimate of the impurity of the node. Based on this, you can visually decide to prune the tree.

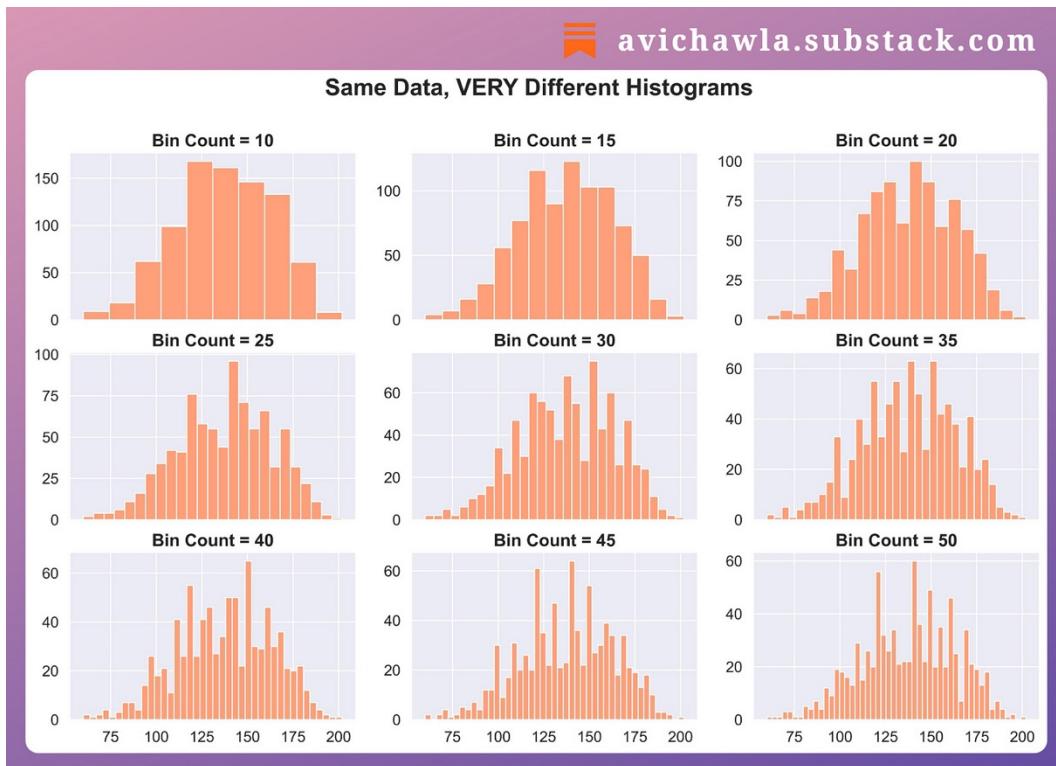
For instance, in the full decision tree shown below, pruning the tree at a depth of two appears to be reasonable.



Once you have obtained a rough estimate for these hyperparameter values, you can train a new decision tree. Next, measure its performance on new data to know if the decision tree is generalizing or not.



# Use Histograms With Caution. They Are Highly Misleading!



Histograms are commonly used for data visualization. But, they can be misleading at times. Here's why.

Histograms divide the data into small bins and represent the frequency of each bin.

Thus, the choice of the number of bins you begin with can significantly impact its shape.

The figure above depicts the histograms obtained on the same data, but by altering the number of bins. Each histogram conveys a different story, even though the underlying data is the same.

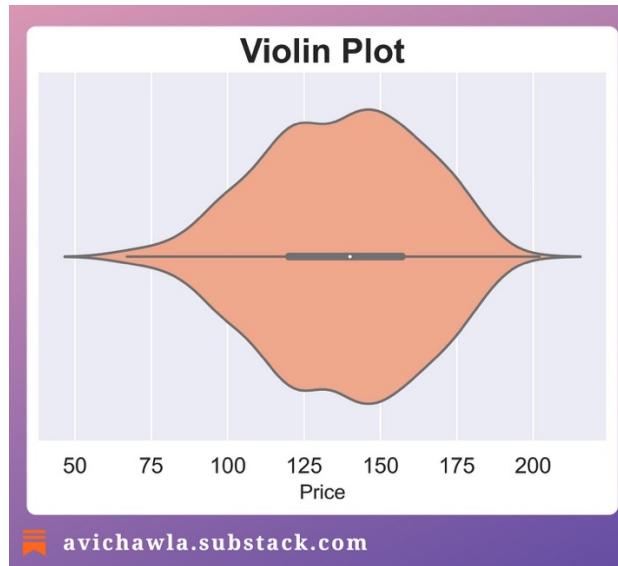
This, at times, can be misleading and may lead you to draw the wrong conclusions.

The takeaway is NOT that histograms should not be used. Instead, look at the underlying distribution too. Here, a violin plot and a KDE plot can help.

**Violin plot**



Similar to box plots, Violin plots also show the distribution of data based on quartiles. However, it also adds a kernel density estimation to display the density of data at different values.



avichawla.substack.com

This provides a more detailed view of the distribution, particularly in areas with higher density.

### KDE plot

KDE plots use a smooth curve to represent the data distribution, without the need for binning, as shown below:



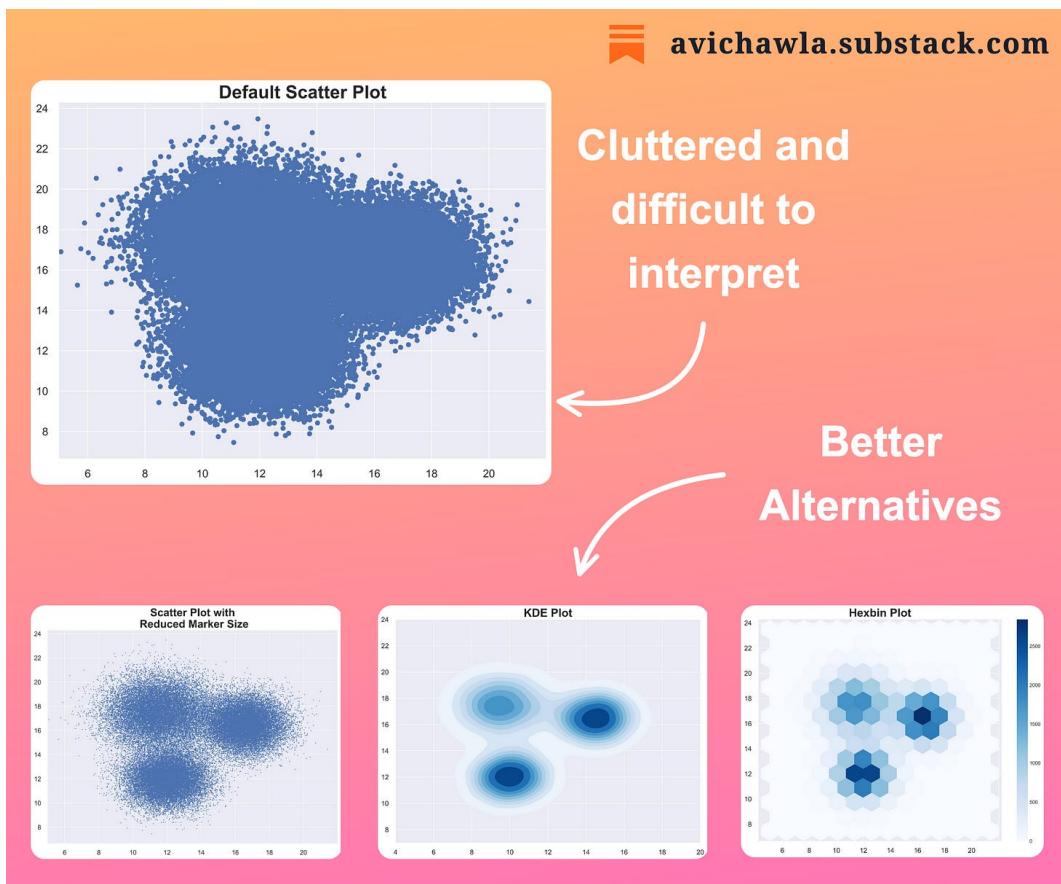
avichawla.substack.com

As a departing note, always remember that whenever you condense a dataset, you run the risk of losing important information.

Thus, be mindful of any limitations (and assumptions) of the visualizations you use. Also, consider using multiple methods to ensure that you are seeing the whole picture.



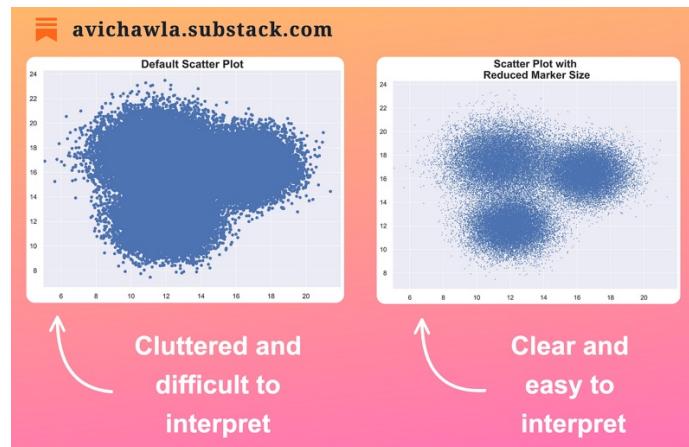
# Three Simple Ways To (Instantly) Make Your Scatter Plots Clutter Free



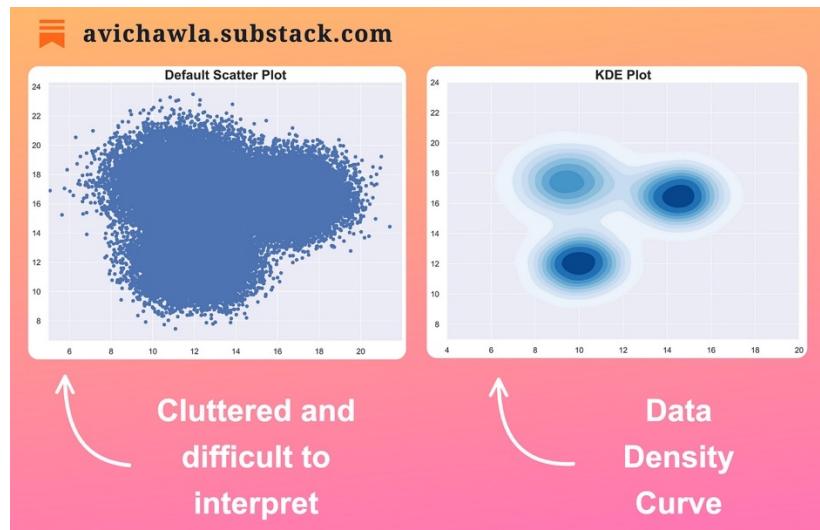
Scatter plots are commonly used in data visualization tasks. But when you have many data points, they often get too dense to interpret.

Here are a few techniques (and alternatives) you can use to make your data more interpretable in such cases.

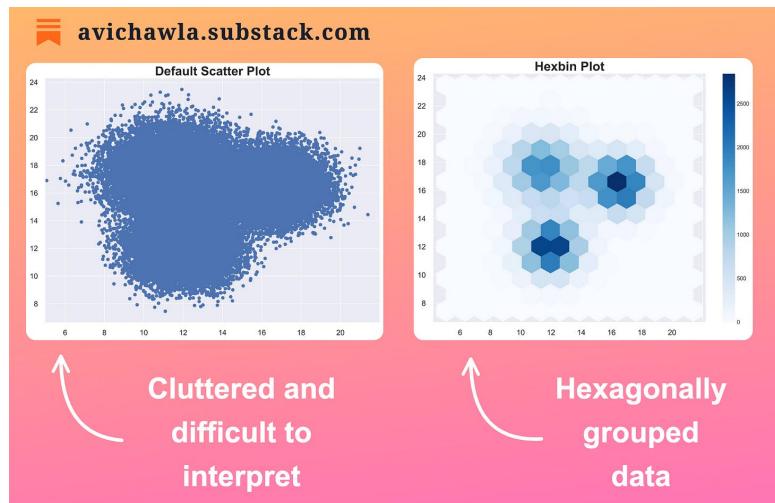
One of the simplest yet effective ways could be to reduce the marker size. This, at times, can instantly offer better clarity over the default plot.



Next, as an alternative to a scatter plot, you can use a density plot, which depicts the data distribution. This makes it easier to identify regions of high and low density, which may not be evident from a scatter plot.

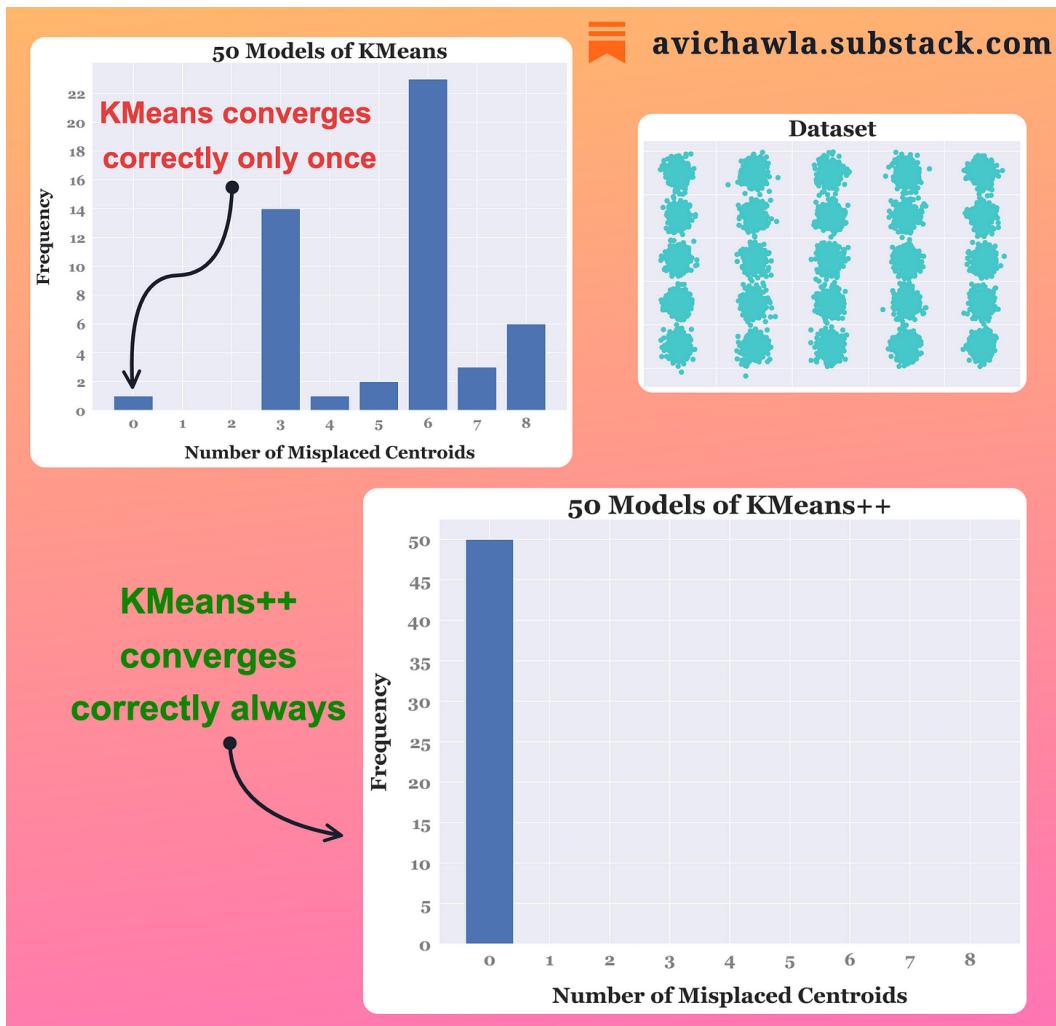


Lastly, another better alternative can be a hexbin plot. It bins the chart into hexagonal regions and assigns a color intensity based on the number of points in that area.





# A (Highly) Important Point to Consider Before You Use KMeans Next Time



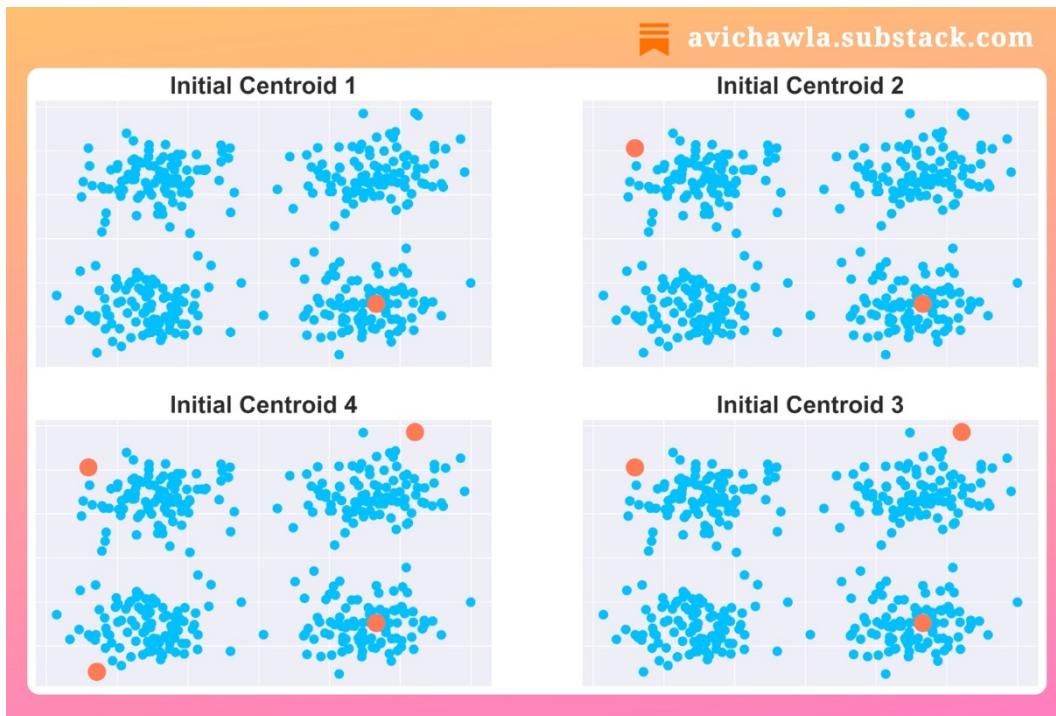
The most important yet often overlooked step of KMeans is its centroid initialization. Here's something to consider before you use it next time.

KMeans selects the initial centroids randomly. As a result, it fails to converge at times. This requires us to repeat clustering several times with different initialization.

Yet, repeated clustering may not guarantee that you will soon end up with the correct clusters. This is especially true when you have many centroids to begin with.

Instead, KMeans++ takes a smarter approach to initialize centroids.

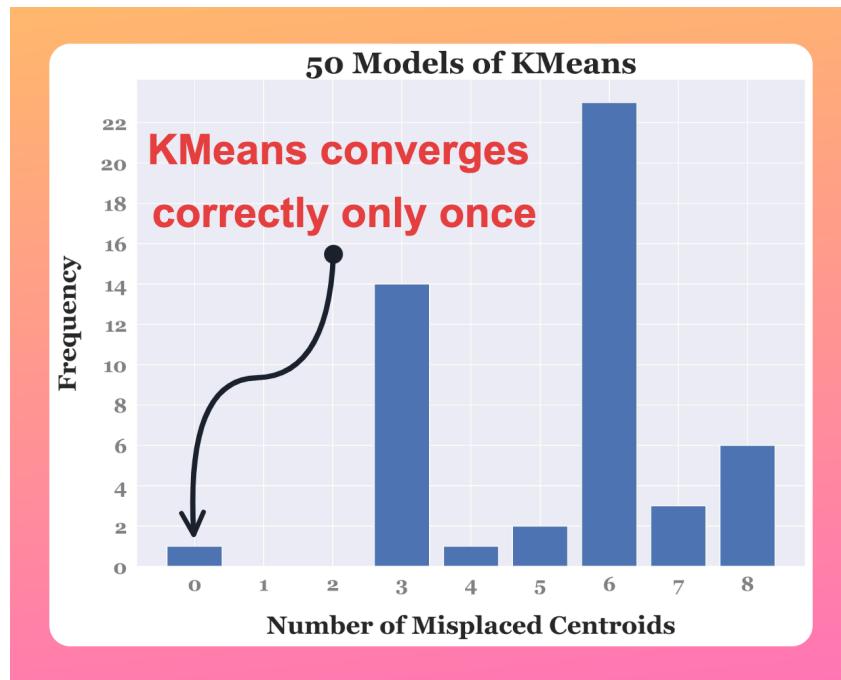
The first centroid is selected randomly. But the next centroid is chosen based on the distance from the first centroid.



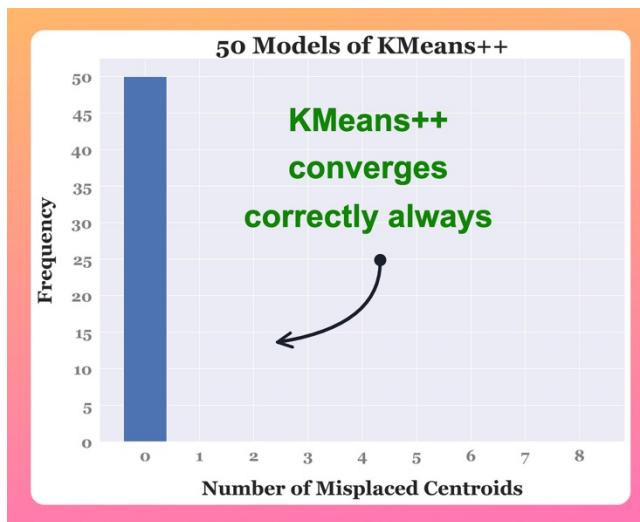
In other words, a point that is away from the first centroid is more likely to be selected as an initial centroid. This way, all the initial centroids are likely to lie in different clusters already, and the algorithm may converge faster and more accurately.

The impact is evident from the bar plots shown below. They depict the frequency of the number of misplaced centroids obtained (analyzed manually) after training 50 different models with KMeans and KMeans++.

On the given dataset, out of the 50 models, KMeans only produced zero misplaced centroids once, which is a success rate of just **2%**.



In contrast, KMeans++ never produced any misplaced centroids.

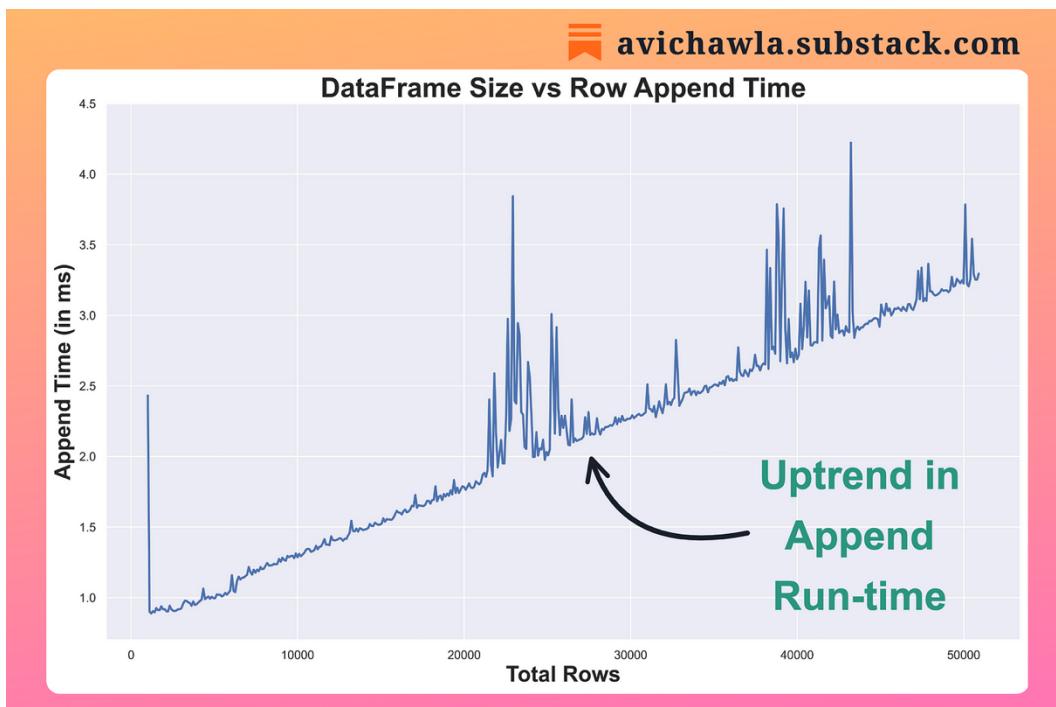


Luckily, if you are using sklearn, you don't need to worry about the initialization step. This is because sklearn, by default, resorts to the KMeans++ approach.

However, if you have a custom implementation, do give it a thought.

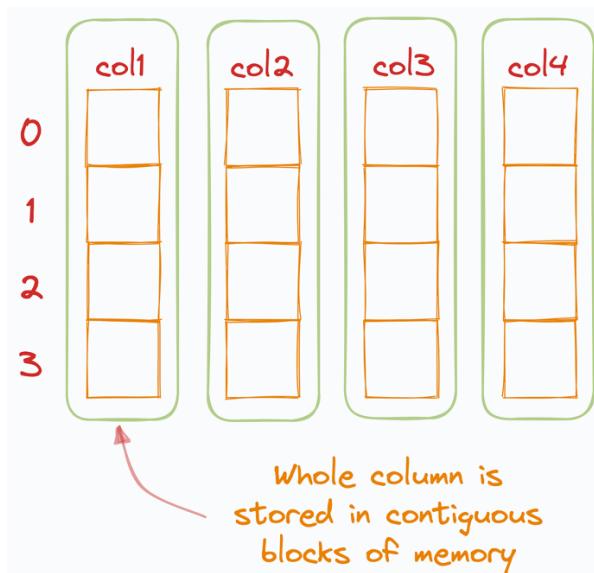


# Why You Should Avoid Appending Rows To A DataFrame



As we append more and more rows to a Pandas DataFrame, the append run-time keeps increasing. Here's why.

A DataFrame is a column-major data structure. Thus, consecutive elements in a column are stored next to each other in memory.





As new rows are added, Pandas always wants to preserve its column-major form.

But while adding new rows, there may not be enough space to accommodate them while also preserving the column-major structure.

In such a case, existing data is moved to a new memory location, where Pandas finds a contiguous block of memory.

Thus, as the size grows, memory reallocation gets more frequent, and the run time keeps increasing.

The reason for spikes in this graph may be because a column taking higher memory was moved to a new location at this point, thereby taking more time to reallocate, or many columns were shifted at once.

### **So what can we do to mitigate this?**

The increase in run-time solely arises because Pandas is trying to maintain its column-major structure.

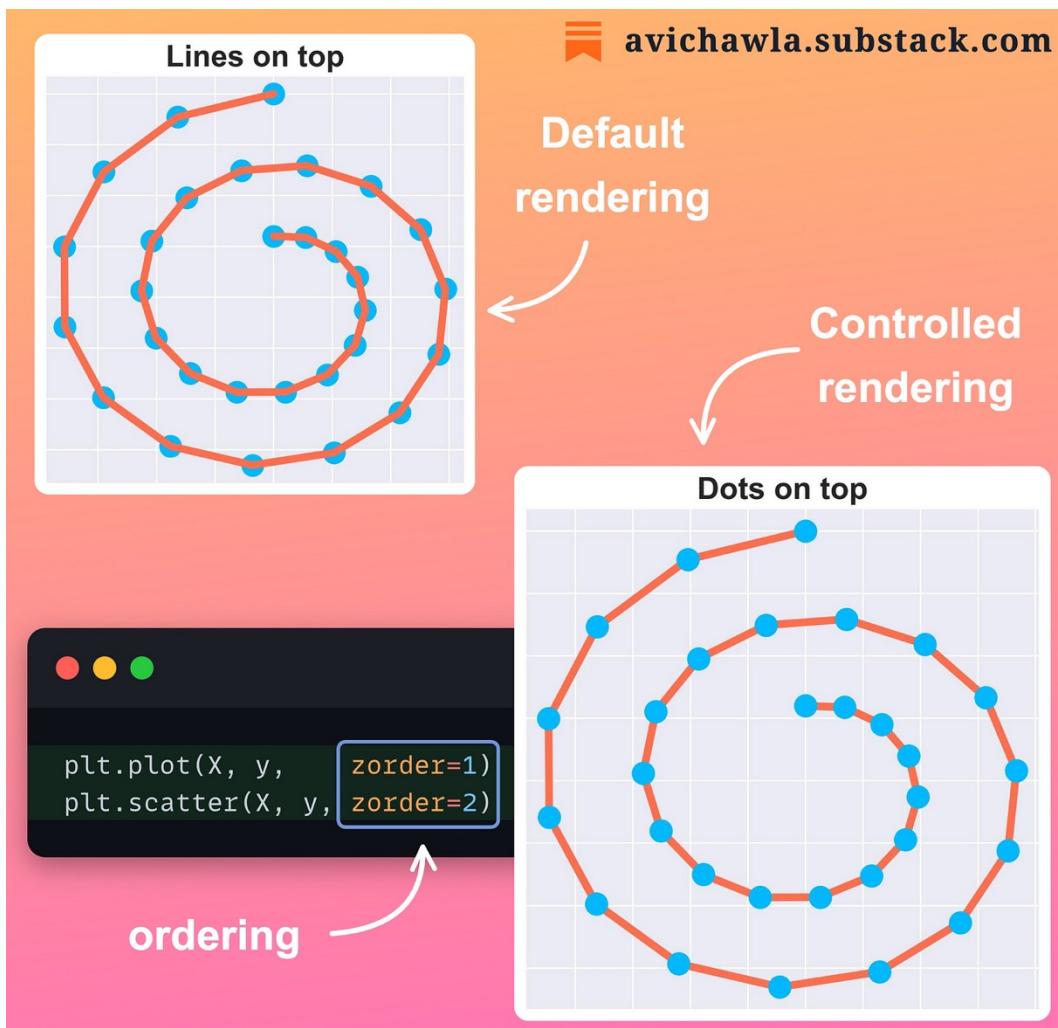
Thus, if you intend to grow a dataframe (row-wise) this frequently, it is better to first convert the dataframe to another data structure, a dictionary or a numpy array, for instance.

Carry out the append operations here, and when you are done, convert it back to a dataframe.

P.S. Adding new columns is not a problem. This is because this operation does not conflict with other columns.



# Matplotlib Has Numerous Hidden Gems. Here's One of Them.



One of the best yet underrated and underutilized potentials of matplotlib is customizability. Here's a pretty interesting thing you can do with it.

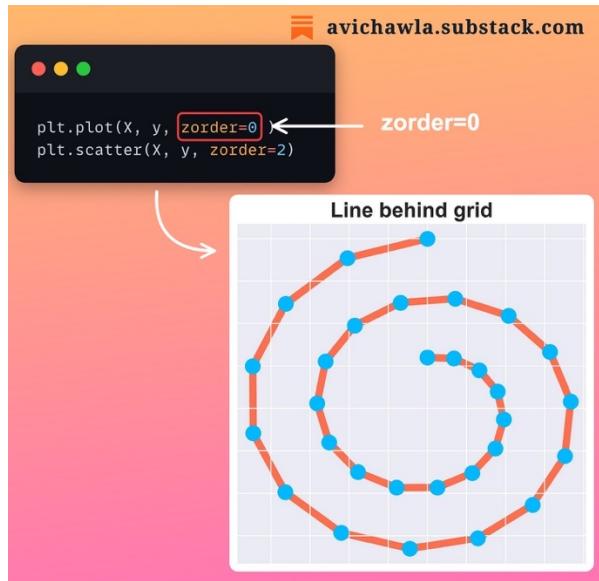
By default, matplotlib renders different types of elements (also called artists), like plots, legend, texts, etc., in a specific order.

But this ordering may not be desirable in all cases, especially when there are overlapping elements in a plot, or the default rendering is hiding some crucial details.

With the `zorder` parameter, you can control this rendering order. As a result, plots with higher `zorder` value appear closer to the viewer and are drawn on top of artists with lower `zorder` values.



Lastly, in the above demonstration, if we specify `zorder=0` for the line plot, we notice that it goes behind the grid lines.



You can find more details about `zorder` here: [Matplotlib docs](#).



# A Counterintuitive Thing About Python Dictionaries

```
>>> my_dict = {  
    1.0 : 'One (float)',  
    1   : 'One (int)',  
    True : 'One (bool)',  
    '1'   : 'One (string)'  
}  
  
>>> my_dict  
{1.0 : 'One (bool)', '1' : 'One (string)'}
```

Added 4 keys

dict only has 2 keys

Despite adding 4 distinct keys to a Python dictionary, can you tell why it only preserves two of them?

Here's why.

In Python, dictionaries find a key based on the equivalence of hash (computed using `hash()`), but not identity (computed using `id()`).

In this case, there's no doubt that `1.0`, `1`, and `True` inherently have different datatypes and are also different objects. This is shown below:



```
>>> id(1.0), id(1), id(True)
(153733, 127473, 493931)

>>> type(1.0), type(1), type(True)
(float, int, bool)
```

Yet, as they share the same hash value, the dictionary considers them as the same keys.

```
>>> hash(1.0), hash(1), hash(True)
(1, 1, 1) ## same hash
```

But did you notice that in the demonstration, the final key is 1.0, while the value corresponds to the key True.

```
>>> my_dict
{1.0: 'One (bool)', '1': 'One (string)'}
```

Annotations:

- A curly arrow points from the text "float key" to the key `1.0`.
- A curly arrow points from the text "value of boolean key" to the value `'One (bool)'`.



This is because, at first, `1.0` is added as a key and its value is '`One (float)`'. Next, while adding the key `1`, python recognizes it as an equivalence of the hash value.

Thus, the value corresponding to `1.0` is overwritten by '`One (int)`', while the key (`1.0`) is kept as is.

Finally, while adding `True`, another hash equivalence is encountered with an existing key of `1.0`. Yet again, the value corresponding to `1.0`, which was updated to '`One (int)`' in the previous step, is overwritten by '`One (bool)`'.

I am sure you may have already guessed why the string key '`1`' is retained.



# Probably The Fastest Way To Execute Your Python Code

```
result = []
for a in range(10000):
    for b in range(10000):
        if (a+b)%11 == 0:
            result.append((a,b))
```

Python

```
$ python big_loop.py
# Run-time: 10.9s
```

100x Faster

Codon

```
$ codon run big_loop.py
# Run-time: 0.11s
```

Many Python programmers are often frustrated with Python's run-time. Here's how you can make your code blazingly fast by changing just one line.

Codon is an open-source, high-performance Python compiler. In contrast to being an interpreter, it compiles your python code to fast machine code.

Thus, post compilation, your code runs at native machine code speed. As a result, typical speedups are often of the order **50x** or more.

According to the official docs, if you know Python, you already know 99% of Codon. There are very minute differences between the two, which you can read here: [Codon docs](#).



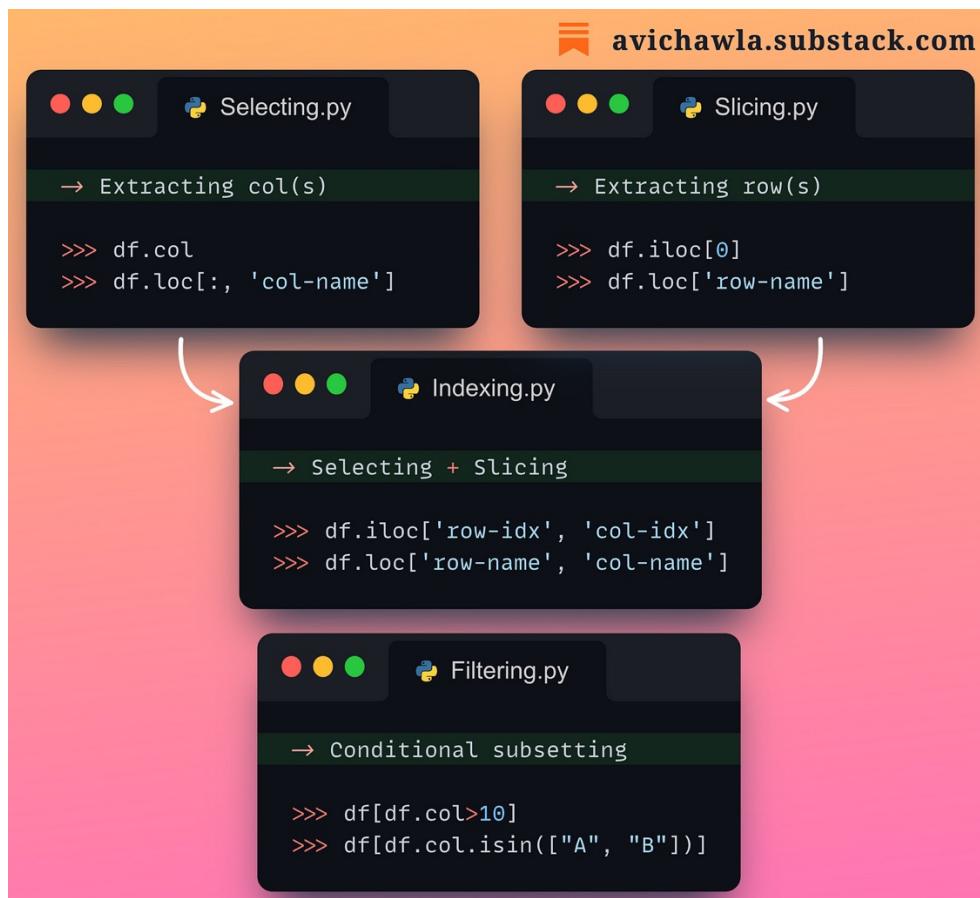
Find some more benchmarking results between Python and Codon below:

The screenshot shows four terminal windows arranged in a 2x2 grid. The top row contains code snippets for `fib.py` and `pi.py`. The bottom row shows the execution times for these scripts using Python and Codon.

Script	Language	Command	Time	Notes
<code>fib.py</code>	Python	<code>\$ python fib.py # N=35</code>	2.53s	
<code>fib.py</code>	Codon	<code>\$ codon run fib.py # N=35</code>	0.04s (~60x Faster)	
<code>fib.py</code>	Python	<code>\$ python fib.py # N=45</code>	296s	
<code>fib.py</code>	Codon	<code>\$ codon run fib.py # N=45</code>	4.89s (~60x Faster)	
<code>pi.py</code>	Python	<code>\$ python pi.py # n_terms=10^8</code>	14.7s	
<code>pi.py</code>	Codon	<code>\$ codon run pi.py # n_terms=10^8</code>	0.35s (~40x Faster)	



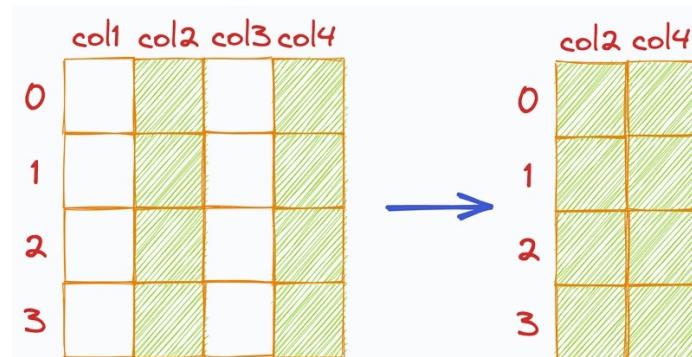
# Are You Sure You Are Using The Correct Pandas Terminologies?



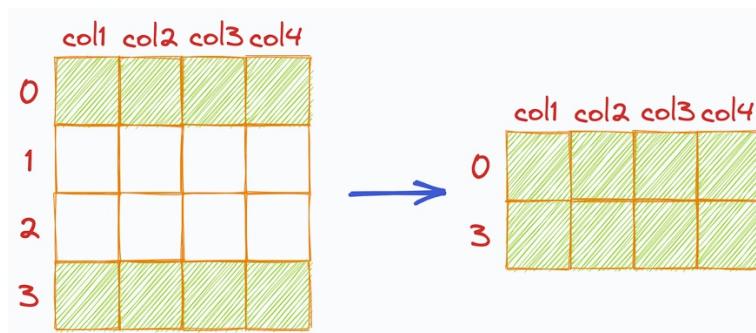
Many Pandas users use the dataframe subsetting terminologies incorrectly. So let's spend a minute to get it straight.

**SUBSETTING** means extracting value(s) from a dataframe. This can be done in four ways:

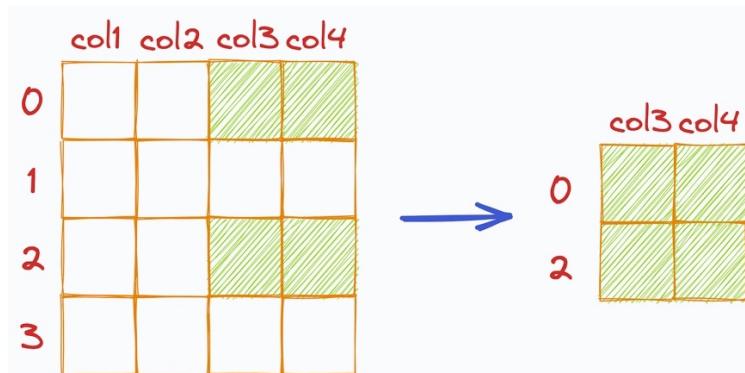
- 1) We call it **SELECTING** when we extract one or more of its **COLUMNS** based on index location or name. The output contains some columns and all rows.



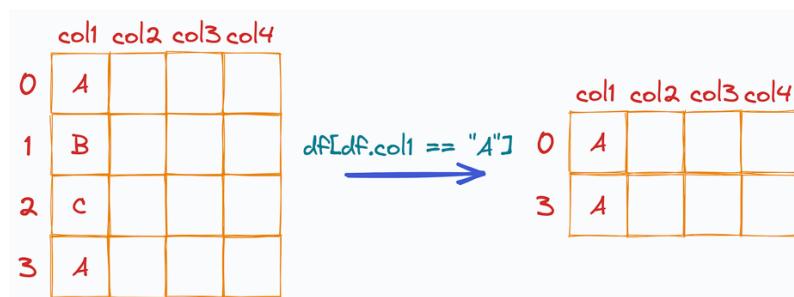
2) We call it **SLICING** when we extract one or more of its **ROWS** based on index location or name. The output contains some rows and all columns.



3) We call it **INDEXING** when we extract both **ROWS** and **COLUMNS** based on index location or name.



4) We call it **FILTERING** when we extract **ROWS** and **COLUMNS** based on conditions.



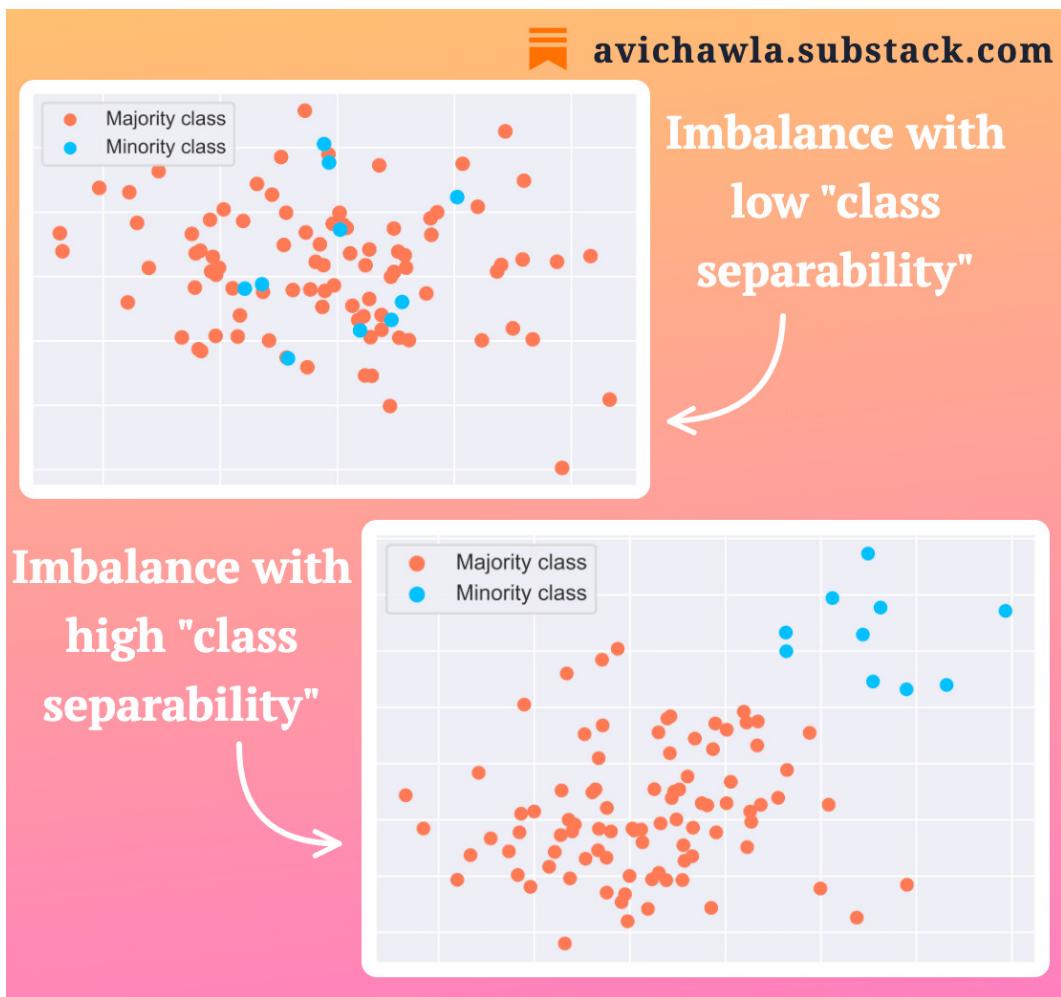


Of course, there are many other ways you can perform these four operations.

Here's a comprehensive Pandas guide I prepared once: [Pandas Map](#). Please refer to the “DF Subset” branch to read about various subsetting methods :)



# Is Class Imbalance Always A Big Problem To Deal With?

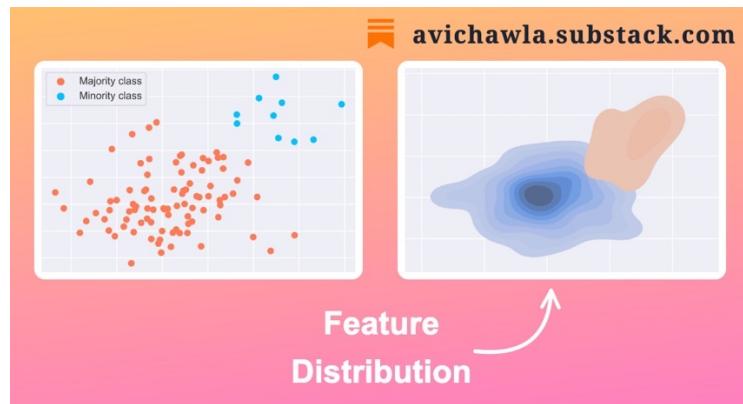


Addressing class imbalance is often a challenge in ML. Yet, it may not always cause a problem. Here's why.

One key factor in determining the impact of imbalance is **class separability**.

As the name suggests, it measures the degree to which two or more classes can be distinguished or separated from each other based on their feature values.

When classes are highly separable, there is little overlap between their feature distributions (as shown below). This makes it easier for a classifier to correctly identify the class of a new instance.

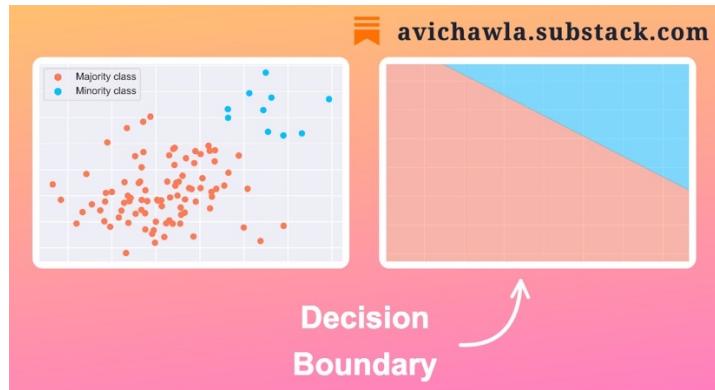


Thus, despite imbalance, even if your data has a high degree of class separability, imbalance may not be a problem per se.

To conclude, consider estimating the class separability before jumping to any sophisticated modeling steps.

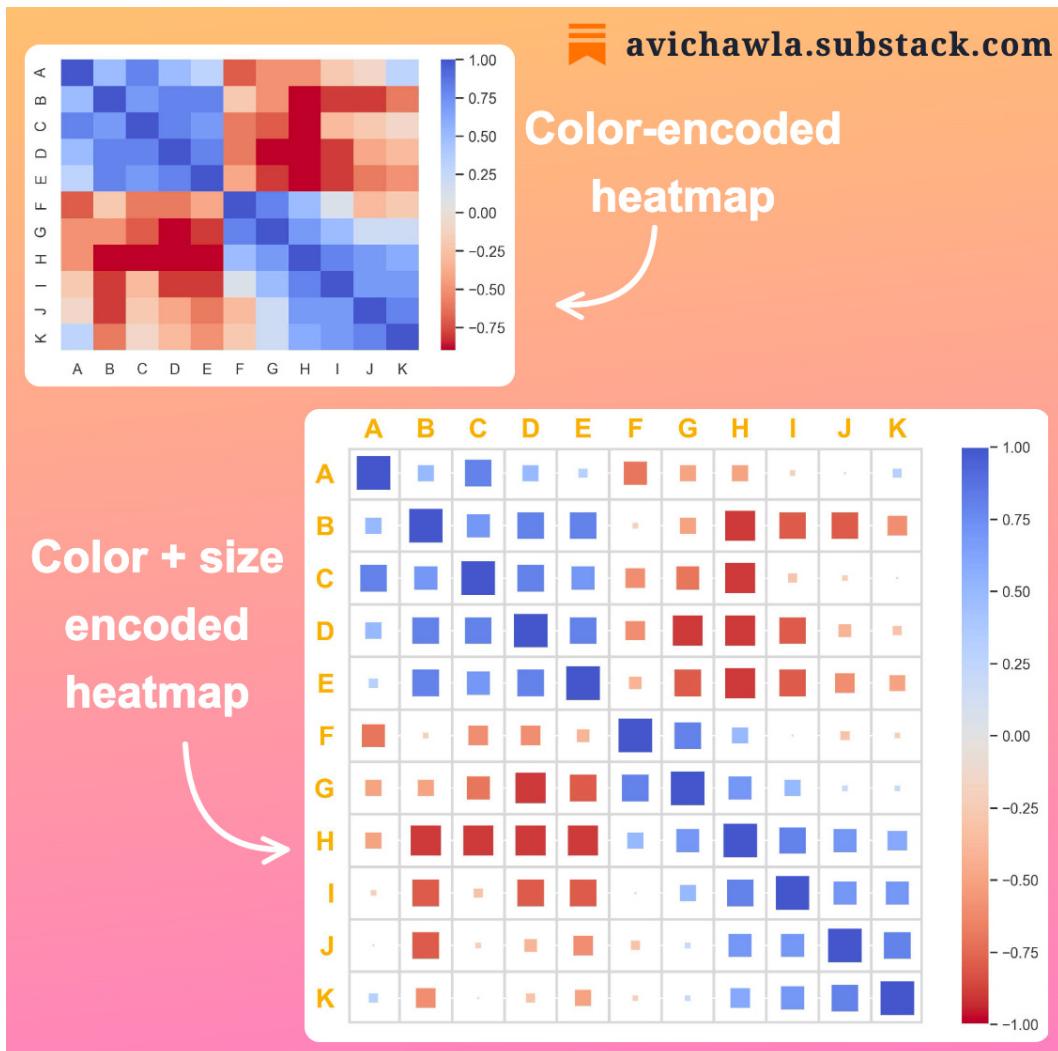
This can be done visually or by evaluating imbalance-specific metrics on simple models.

The figure below depicts the decision boundary learned by a logistic regression model on the class-separable dataset.





# A Simple Trick That Will Make Heatmaps More Elegant



Heatmaps often make data analysis much easier. Yet, they can be further enriched with a simple modification.

A traditional heatmap represents the values using a color scale. Yet, mapping the cell color to numbers is still challenging.

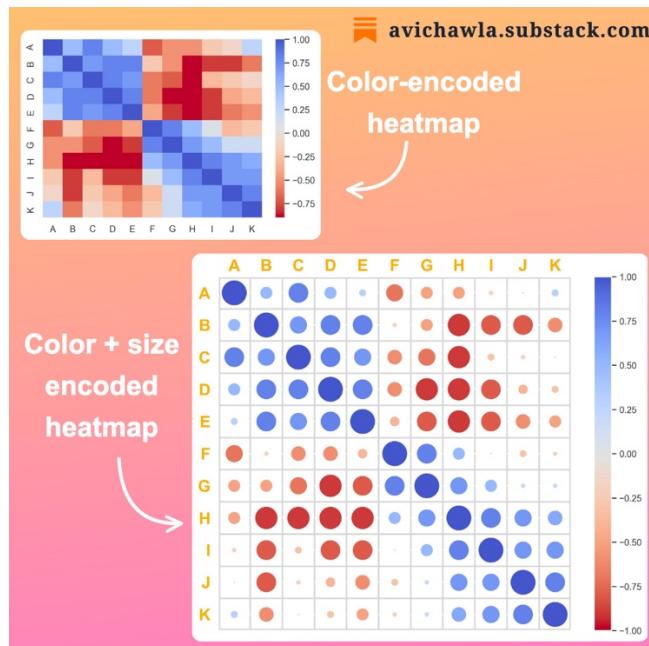
Embedding a size component can be extremely helpful in such cases. In essence, the bigger the size, the higher the absolute value.

This is especially useful to make heatmaps cleaner, as many values nearer to zero will immediately shrink.

In fact, you can represent the size with any other shape. Below, I created the same heatmap using a circle instead:



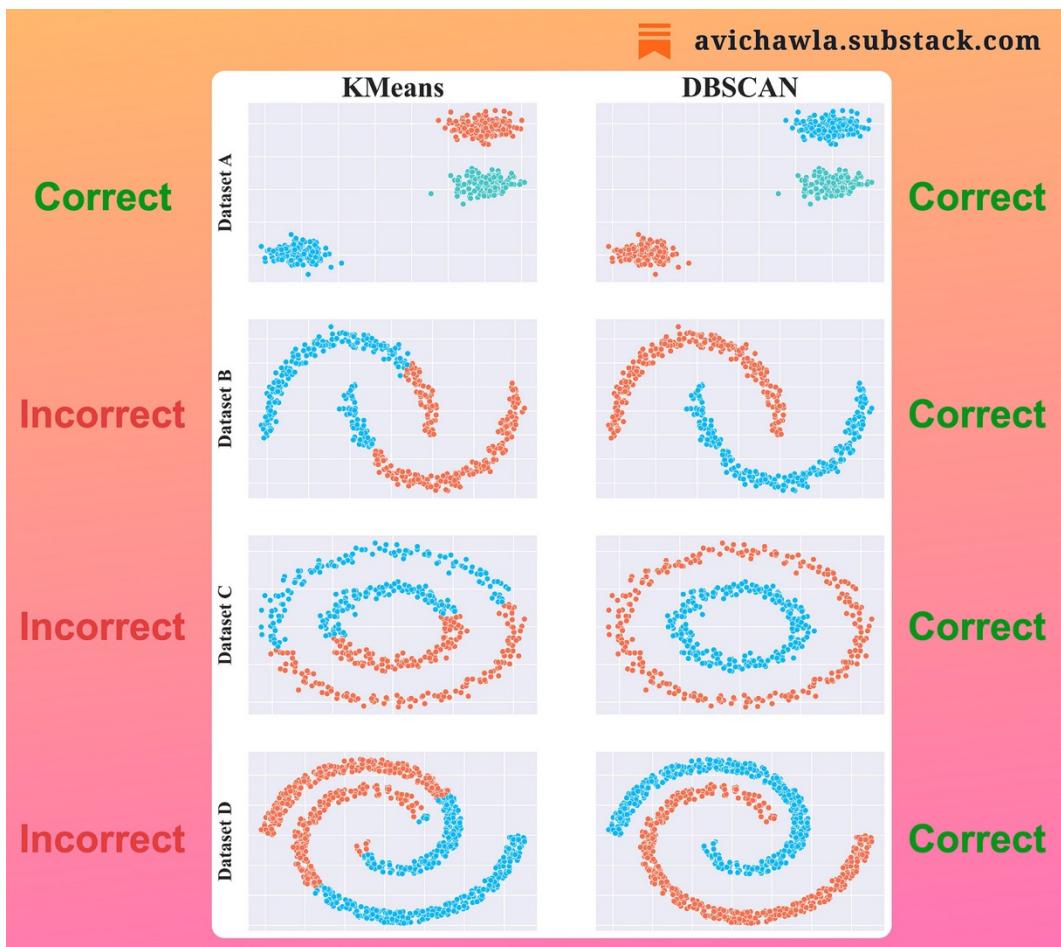
avichawla.substack.com



Find the code for this post here: [GitHub](#).



# A Visual Comparison Between Locality and Density-based Clustering



The utility of KMeans is limited to datasets with spherical clusters. Thus, any variation is likely to produce incorrect clustering.

Density-based clustering algorithms, such as DBSCAN, can be a better alternative in such cases.

They cluster data points based on density, making them robust to datasets of varying shapes and sizes.

The image depicts a comparison of KMeans vs. DBSCAN on multiple datasets.

As shown, KMeans only works well when the dataset has spherical clusters. But in all other cases, it fails to produce correct clusters.

Find more here: [Sklearn Guide](#).

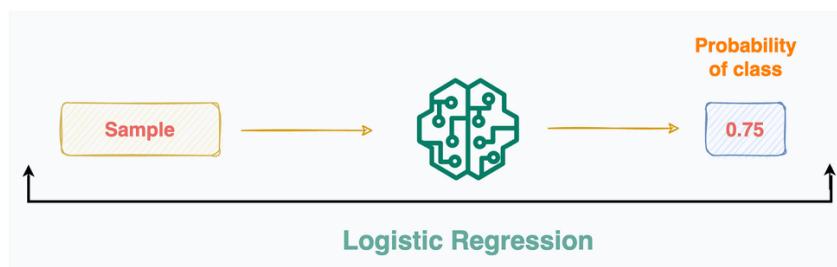


# Why Don't We Call It Logistic Classification Instead?

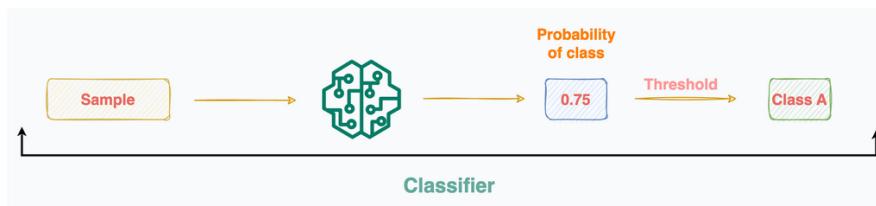


Have you ever wondered why logistic regression is called "regression" when we only use it for classification tasks? Why not call it "logistic classification" instead? Here's why.

Most of us interpret logistic regression as a classification algorithm. However, it is a regression algorithm by nature. This is because it predicts a continuous outcome, which is the probability of a class.



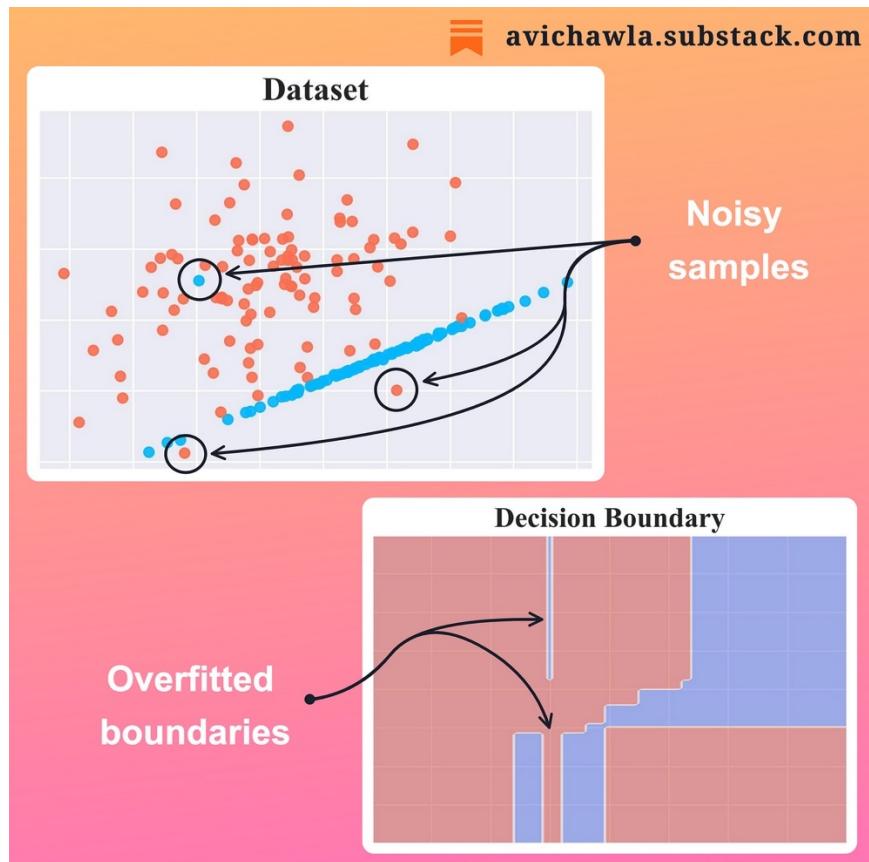
It is only when we apply those thresholds and change the interpretation of its output that the whole pipeline becomes a classifier.



Yet, intrinsically, it is never the algorithm performing the classification. The algorithm always adheres to regression. Instead, it is that extra step of applying probability thresholds that classifies a sample.



# A Typical Thing About Decision Trees Which Many Often Ignore



Although decision trees are simple and intuitive, they always need a bit of extra caution. Here's what you should always remember while training them.

In sklearn's implementation, by default, a decision tree is allowed to grow until all leaves are pure. This leads to overfitting as the model attempts to classify every sample in the training set.

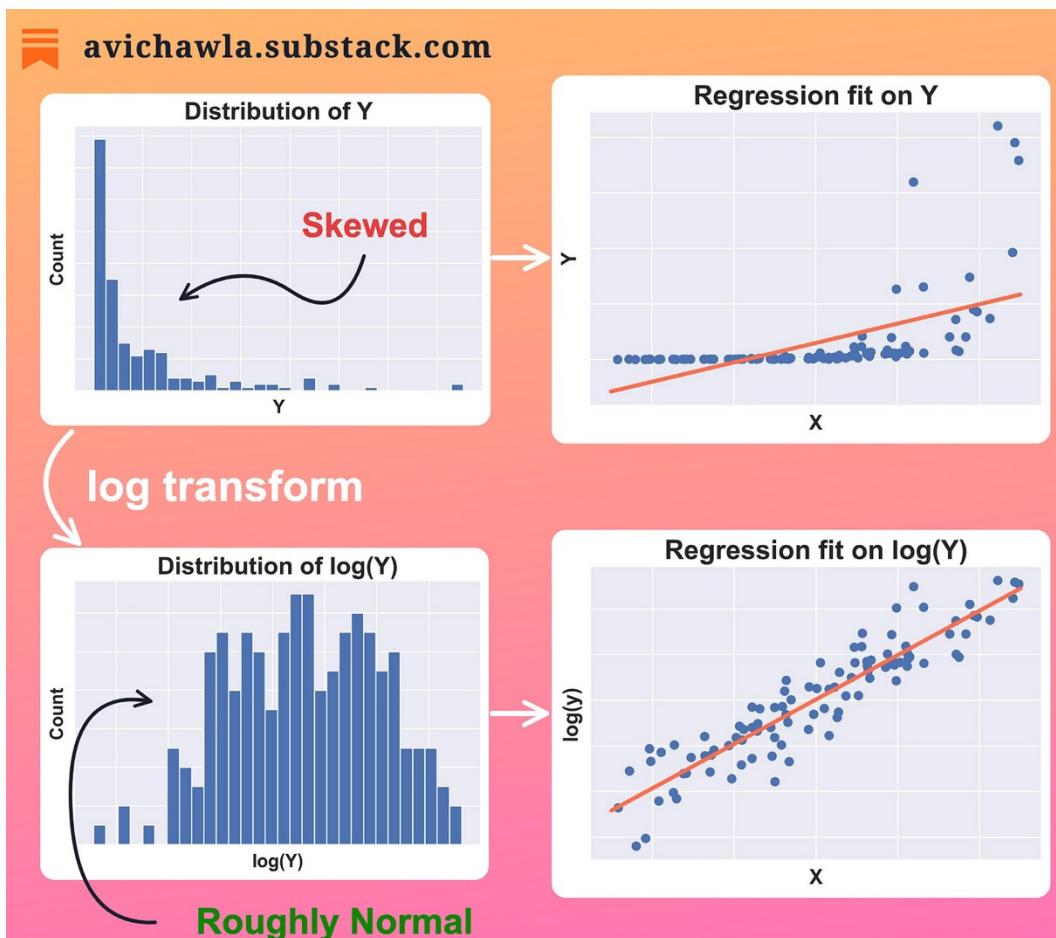
There are various techniques to avoid this, such as pruning and ensembling. Also, make sure that you tune hyperparameters if you use sklearn's implementation.

This was a gentle reminder as many of us often tend to use sklearn's implementations in their default configuration.

It is always a good practice to know what a default implementation is hiding underneath.



# Always Validate Your Output Variable Before Using Linear Regression



The effectiveness of a linear regression model largely depends on how well our data satisfies the algorithm's underlying assumptions.

Linear regression inherently assumes that the residuals (actual-prediction) follow a normal distribution. One way this assumption may get violated is when your output is skewed.

As a result, it will produce an incorrect regression fit.

But the good thing is that it can be corrected. One common way to make the output symmetric before fitting a model is to apply a log transform.

It removes the skewness by evenly spreading out the data, making it look somewhat normal.

One thing to note is that if the output has negative values, a log transform will raise an error. In such cases, one can apply translation transformation first on the output, followed by the log.



# A Counterintuitive Fact About Python Functions

```
# Define a function
>>> def my_func(): pass

# 1) Verify the type of function object
>>> type(my_func)
<class 'function'>

# 2) Add new attributes to function object
>>> my_func.my_attr = 'new_attribute'
>>> my_func.my_attr
'new_attribute'

# 3) Pass as an argument to other functions
>>> def new_func(f): pass
>>> new_func(my_func)

# 4) Access instance-level attributes/methods
>>> my_func.__name__
'my_func'
>>> my_func.__dict__
{'my_attr': 'new_attribute'}
```

Everything in python is an object instantiated from some class. This also includes functions, but accepting this fact often feels counterintuitive at first.

Here are a few ways to verify that python functions are indeed objects.

The friction typically arises due to one's acquaintance with other programming languages like C++ and Java, which work very differently.

However, python is purely an object-oriented programming (OOP) language. You are always using OOP, probably without even realizing it.



# Why Is It Important To Shuffle Your Dataset Before Training An ML Model



ML models may fail to converge for many reasons. Here's one of them which many folks often overlook.

If your data is ordered by labels, this could negatively impact the model's convergence and accuracy. This is a mistake that can typically go unnoticed.

In the above demonstration, I trained two neural nets on the same data. Both networks had the same initial weights, learning rate, and other settings.

However, in one of them, the data was ordered by labels, while in another, it was randomly shuffled.

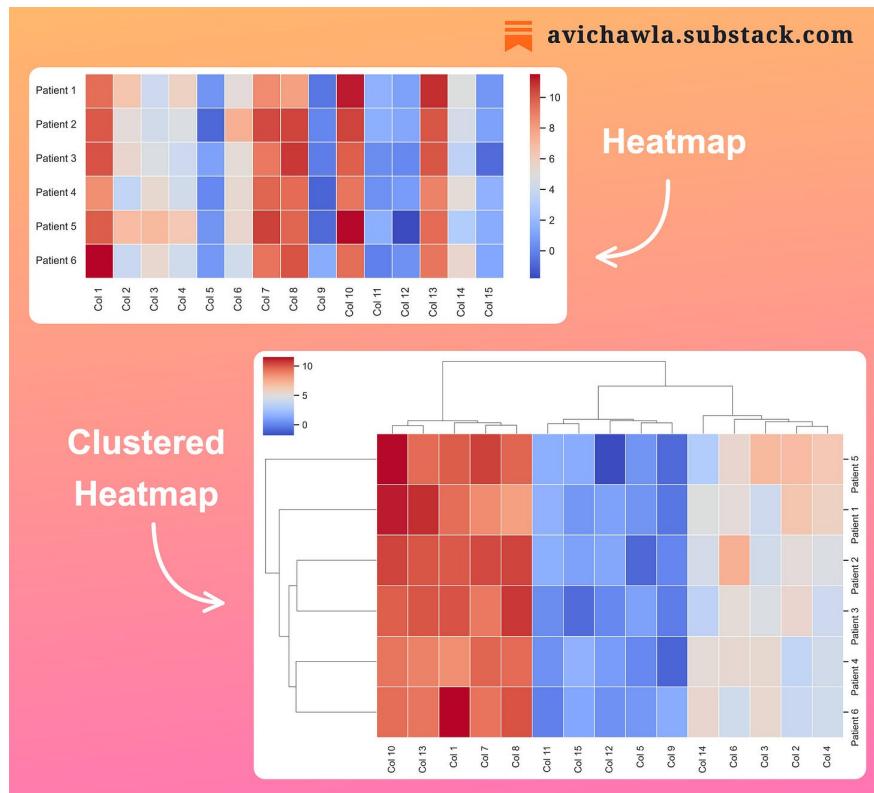
As shown, the model receiving a label-ordered dataset fails to converge. However, shuffling the dataset allows the network to learn from a more representative data sample in each batch. This leads to better generalization and performance.

In general, it's a good practice to shuffle the dataset before training. This prevents the model from identifying any label-specific yet non-existing patterns.

In fact, it is also recommended to alter batch-specific data in every epoch.



# The Limitations Of Heatmap That Are Slowing Down Your Data Analysis



Heatmaps often make data analysis much easier. Yet, they do have some limitations.

A traditional heatmap does not group rows (and features). Instead, its orientation is the same as the input. This makes it difficult to visually determine the similarity between rows (and features).

Clustered heatmaps can be a better choice in such cases. It clusters the rows and features together to help you make better sense of the data.

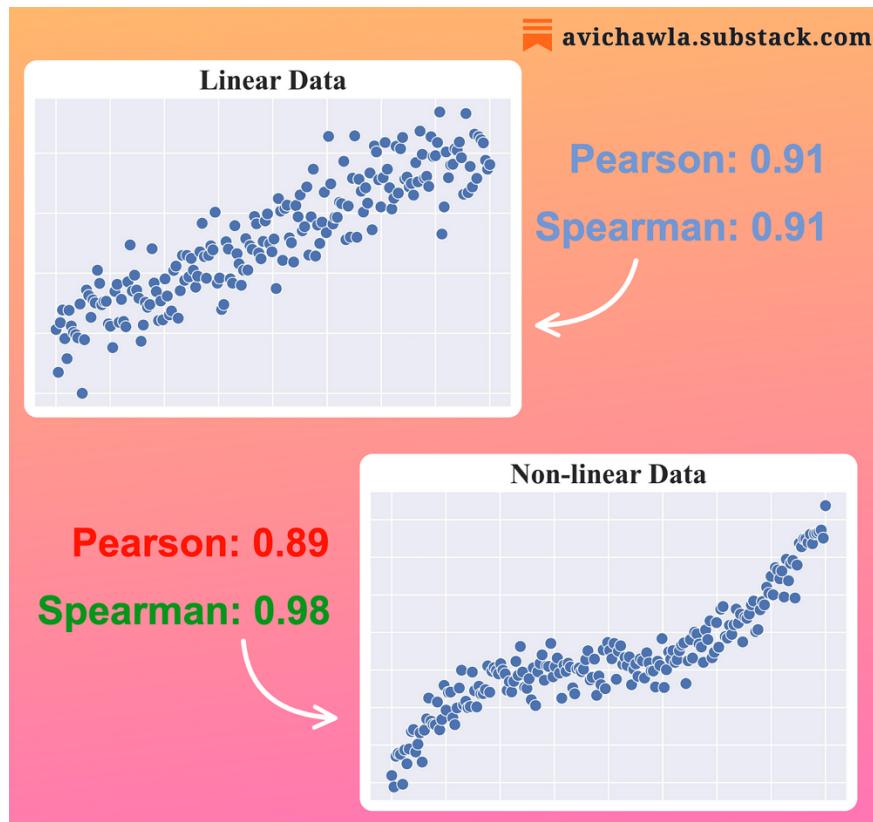
They can be especially useful when dealing with large datasets. While a traditional heatmap will be visually daunting to look at.

However, the groups in a clustered heatmap make it easier to visualize similarities and identify which rows (and features) go with one another.

To create a clustered heatmap, you can use the `sns.clustermap()` method from Seaborn. More info here: [Seaborn docs](#).



# The Limitation Of Pearson Correlation Which Many Often Ignore



Pearson correlation is commonly used to determine the association between two continuous variables. But many often ignore its assumption.

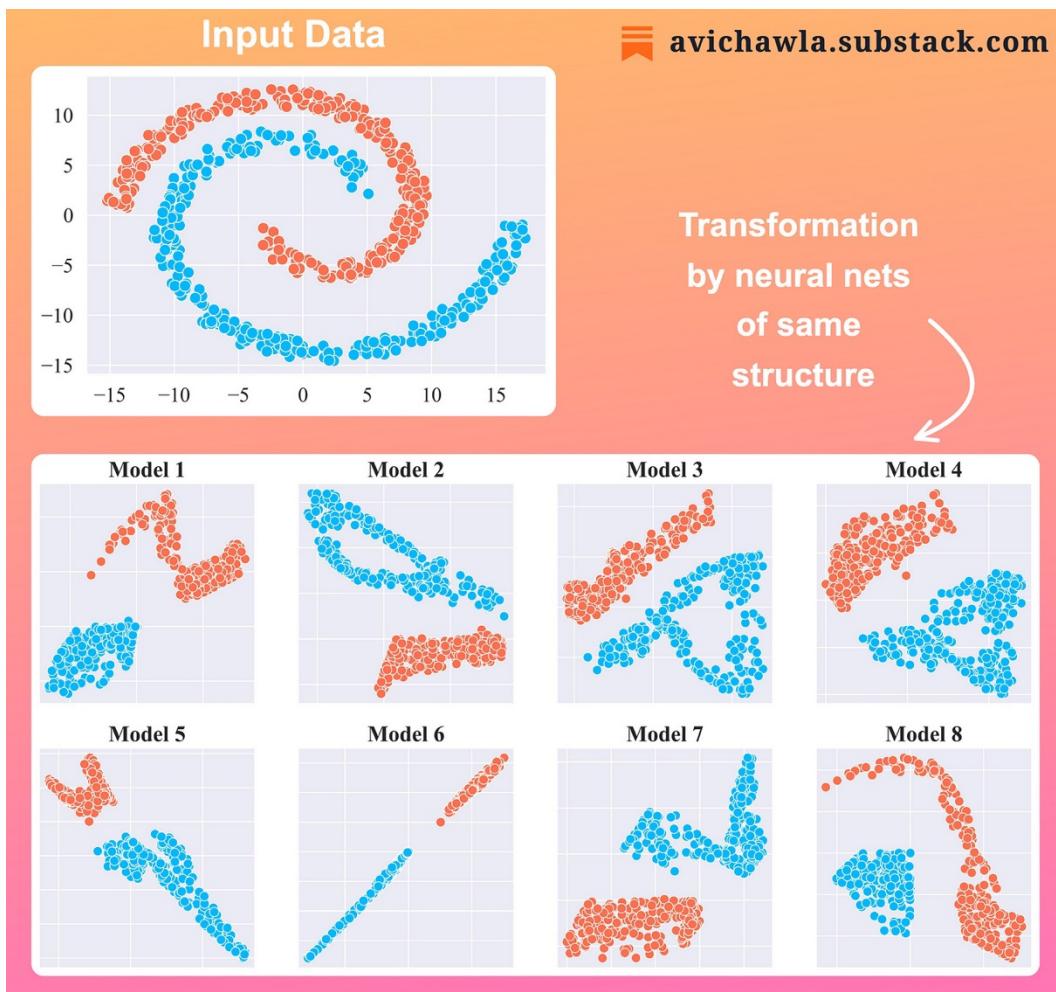
Pearson correlation primarily measures the LINEAR relationship between two variables. As a result, even if two variables have a non-linear but monotonic relationship, Pearson will penalize that.

One great alternative is the Spearman correlation. It primarily assesses the monotonicity between two variables, which may be linear or non-linear.

What's more, Spearman correlation is also useful in situations when your data is ranked or ordinal.



# Why Are We Typically Advised To Set Seeds for Random Generators?



From time to time, we advised to set seeds for random numbers before training an ML model. Here's why.

The weight initialization of a model is done randomly. Thus, any repeated experiment never generates the same set of numbers. This can hinder the reproducibility of your model.

As shown above, the same input data gets transformed in many ways by different neural networks of the same structure.

Thus, before training any model, always ensure that you set seeds so that your experiment is reproducible later.



# An Underrated Technique To Improve Your Data Visualizations



At times, ensuring that your plot conveys the right message may require you to provide additional context. Yet, augmenting extra plots may clutter your whole visualization.

One great way to provide extra info is by adding text annotations to a plot.

In matplotlib, you can use **annotate()**. It adds explanatory texts to your plot, which lets you guide a viewer's attention to specific areas and aid their understanding.

Find more info here: [Matplotlib docs](#).



# A No-Code Tool to Create Charts and Pivot Tables in Jupyter

The screenshot shows a Jupyter notebook titled "notebook.ipynb". The code cell contains:

```
from pivottablejs import pivot_ui  
pivot_ui(df)
```

Below the code cell is a screenshot of the PivotTableJS interface. The interface has a sidebar with dropdown menus for "Table", "Count", and "Employment\_Status". The main area shows two columns: "Name" and "Employee\_City". The "Employment\_Status" column is expanded, showing a table with data for various cities. The table includes columns for "Employment\_Status" (Full Time, Intern, Totals) and rows for cities like Aliciafort, Kristaburgh, New Cindychester, etc.

Employee_City	Employment_Status	Full Time	Intern	Totals
Aliciafort		89	24	113
Kristaburgh		77	20	97
New Cindychester		82	24	106
New Russellton		73	20	93
North Melissafurt		62	16	78
Ricardomouth		81	25	106
Wardfort		82	14	96
West Jamesview		94	26	120
Whitakerbury		66	21	87
Whiteside		85	19	104
Totals		791	209	1,000

Here's a quick and easy way to create pivot tables, charts, and group data without writing any code.

PivotTableJS is a drag-n-drop tool for creating pivot tables and interactive charts in Jupyter. What's more, you can also augment pivot tables with heatmaps for enhanced analysis.

Find more info here: [PivotTableJS](#).

**Watch a video version of this post for enhanced understanding: [Video](#).**



# If You Are Not Able To Code A Vectorized Approach, Try This.

df.shape	avichawla.substack.com
(100000, 9)	
<b>1) iterrows()</b>	
<pre>%timeit [my_func(row) for index, row in df.iterrows()]</pre>	
2.63 s ± 7.55 ms per loop	<b>Slowest</b>
<b>2) apply()</b>	
<pre>%timeit df.apply(my_func, axis = 1)</pre>	
923 ms ± 6 ms per loop	<b>Slow</b>
<b>3) itertuples()</b>	
<pre>%timeit [my_func(row) for row in df.itertuples()]</pre>	
87.3 ms ± 486 µs per loop	<b>Fast</b>
<b>4) to_numpy()</b>	
<pre>%timeit np_arr = df.to_numpy(); [my_func(row) for row in np_arr]</pre>	
32.9 ms ± 240 µs per loop	<b>Fastest</b>

Although we should never iterate over a dataframe and prefer vectorized code, what if we are not able to come up with a vectorized solution?

In my yesterday's post on why iterating a dataframe is costly, someone posed a pretty genuine question. They asked: "*Let's just say you are forced to iterate. What will be the best way to do so?*"

Firstly, understand that the primary reason behind the slowness of iteration is due to the way a dataframe is stored in memory. (If you wish to recap this, read yesterday's post [here](#).)

Being a column-major data structure, retrieving its rows requires accessing non-contiguous blocks of memory. This increases the run-time drastically.



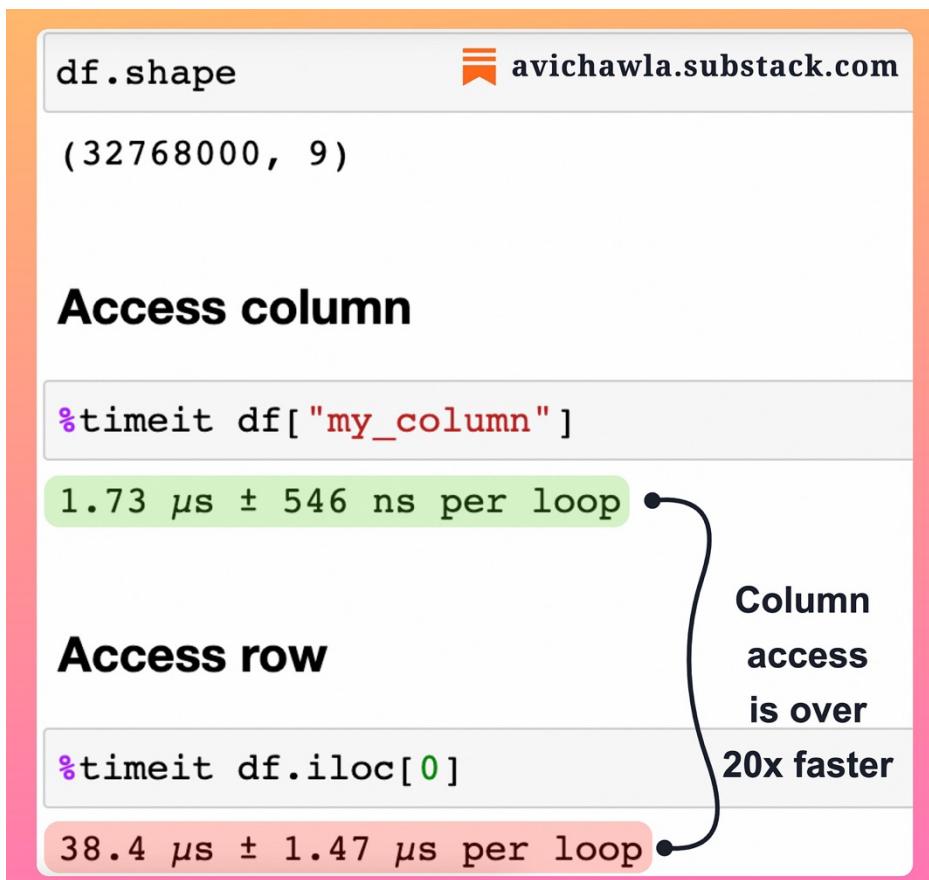
Yet, if you wish to perform only row-based operations, a quick fix is to convert the dataframe to a NumPy array.

NumPy is faster here because, by default, it stores data in a row-major manner. Thus, its rows are retrieved by accessing contiguous blocks of memory, making it efficient over iterating a dataframe.

That being said, do note that the best way is to write vectorized code always. Use the Pandas-to-NumPy approach only when you are truly struggling with writing vectorized code.



# Why Are We Typically Advised To Never Iterate Over A DataFrame?



From time to time, we are advised to avoid iterating on a Pandas DataFrame. But what is the exact reason behind this? Let me explain.

A DataFrame is a column-major data structure. Thus, consecutive elements in a column are stored next to each other in memory.

As processors are efficient with contiguous blocks of memory, retrieving a column is much faster than a row.

But while iterating, as each row is retrieved by accessing non-contiguous blocks of memory, the run-time increases drastically.

In the image above, retrieving over 32M elements of a column was still over **20x faster** than fetching just nine elements stored in a row.



# Manipulating Mutable Objects In Python Can Get Confusing At Times

```
Method1.py
1 # 1) Define list
2 >>> a = [1,2,3]
3
4 # 2) Assign b to a
5 >>> b = a
6
7 # 3) Modify a
8 >>> a = a + [4,5]
9
10 # 4) Print a
11 >>> a
12 [1, 2, 3, 4, 5] # Modified
13
14 # 5) Print b
15 >>> b
16 [1, 2, 3] # Unchanged

Method2.py
1 # 1) Define list
2 >>> a = [1,2,3]
3
4 # 2) Assign b to a
5 >>> b = a
6
7 # 3) Modify a
8 >>> a += [4,5]
9
10 # 4) Print a
11 >>> a
12 [1, 2, 3, 4, 5] # Modified
13
14 # 5) Print b
15 >>> b
16 [1, 2, 3, 4, 5] # Modified
```

Did you know that with mutable objects, “`a +=`” and “`a = a +`” work differently in Python? Here's why.

Let's consider a list, for instance.

When we use the `=` operator, Python creates a new object in memory and assigns it to the variable.

Thus, all the other variables still reference the previous memory location, which was never updated. This is shown in `Method1.py` above.

But with the `+=` operator, changes are enforced in-place. This means that Python does not create a new object and the same memory location is updated.

Thus, changes are visible through all other variables that reference the same location. This is shown in `Method2.py` above.

We can also verify this by comparing the `id()` pre-assignment and post-assignment.



The image shows two code editors side-by-side. The top editor is titled 'Method1.py' and contains the following code:

```
1 # 1) Check ID
2 >>> id(a), id(b)
3 (12345, 12345)
4
5 # 2) Modify a
6 >>> a = a + [4,5]
7
8 # 3) Check ID
9 >>> id(a), id(b)
10 (98765, 12345)
```

The bottom editor is titled 'Method2.py' and contains the following code:

```
1 # 1) Check ID
2 >>> id(a), id(b)
3 (12345, 12345)
4
5 # 2) Modify a
6 >>> a += [4,5]
7
8 # 3) Check ID
9 >>> id(a), id(b)
10 (12345, 12345)
```

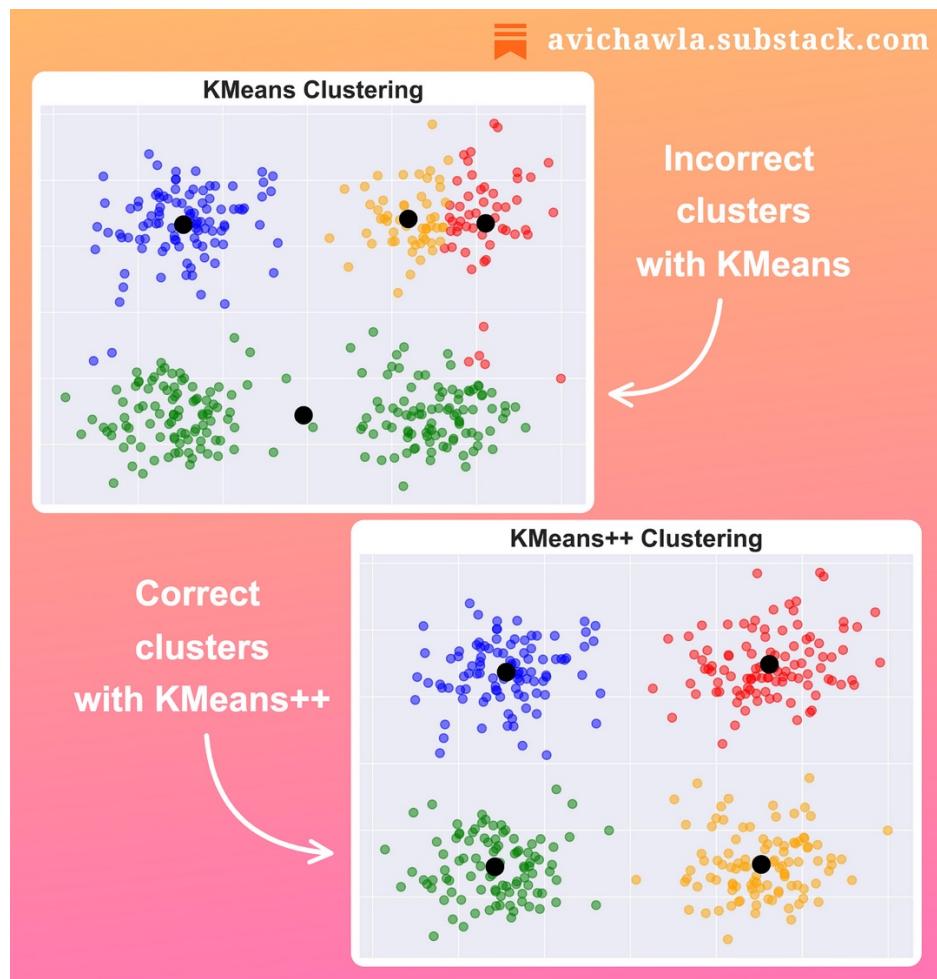
Annotations on the left side of the image explain the behavior of each method:

- A bracket labeled "id(a) changed" points to the output of the first print statement in Method1.py.
- A bracket labeled "id(a) unchanged" points to the output of the first print statement in Method2.py.

With “`a = a +`”, the `id` gets changed, indicating that Python created a new object. However, with “`a +=`”, `id` stays the same. This indicates that the same memory location was updated.



# This Small Tweak Can Significantly Boost The Run-time of KMeans



KMeans is a popular but high-run-time clustering algorithm. Here's how a small tweak can significantly improve its run time.

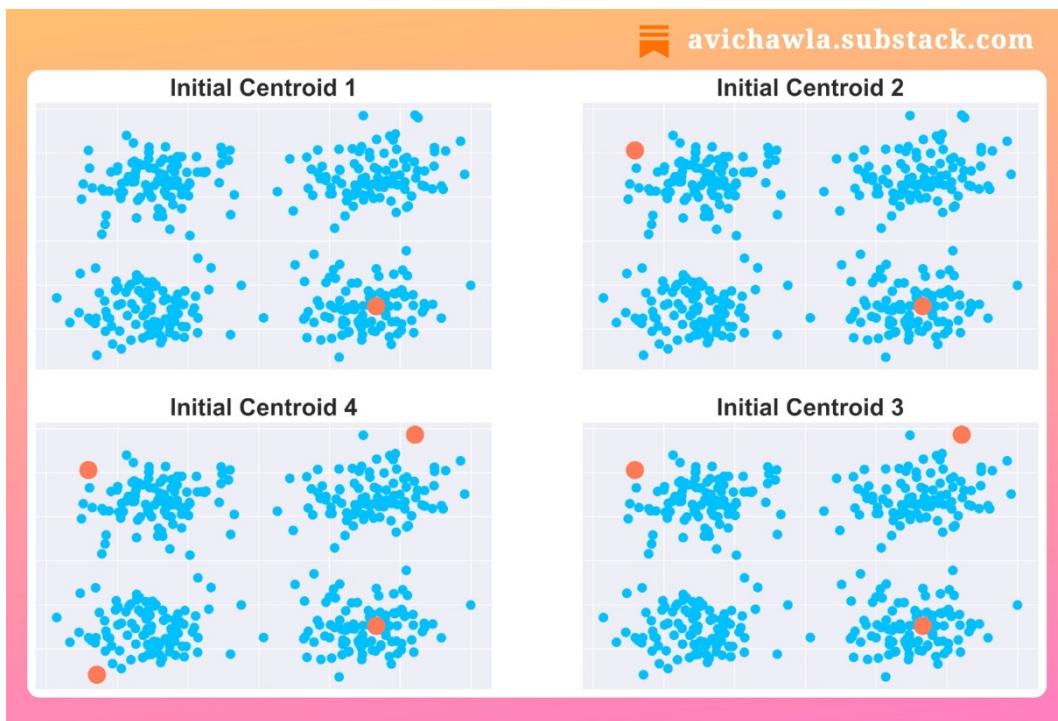
KMeans selects the initial centroids randomly. As a result, it fails to converge at times. This requires us to repeat clustering several times with different initialization.

Instead, KMeans++ takes a smarter approach to initialize centroids. The first centroid is selected randomly. But the next centroid is chosen based on the distance from the first centroid.

In other words, a point that is away from the first centroid is more likely to be selected as an initial centroid. This way, all the initial centroids are likely to lie in different clusters already and the algorithm may converge faster.



The illustration below shows the centroid initialization of KMeans++:





# Most Python Programmers Don't Know This About Python OOP

```
class Point2D:
    def __new__(cls, x, y):
        if isinstance(x, int) and isinstance(y, int):
            # Allocate memory and return a new object
            # only when the if-condition is True
            print("Creating Object!")
            return super().__new__(cls) # Return new object
        else:
            raise TypeError("x and y must be integers")

    def __init__(self, x, y):
        self.x = x
        self.y = y
        print("Object Initialized!")

>>> p1 = Point2D(1,2)
"Creating Object!"    # from __new__() method
"Object Initialized!" # from __init__() method

>>> p2 = Point2D(1.5, 2.5)
TypeError: x and y must be integers
```

Most python programmers misunderstand the `__init__()` method. They think that it creates a new object. But that is not true.

When we create an object, it is not the `__init__()` method that allocates memory to it. As the name suggests, `__init__()` only assigns value to an object's attributes.

Instead, Python invokes the `__new__()` method first to create a new object and allocate memory to it. But how is that useful, you may wonder? There are many reasons.

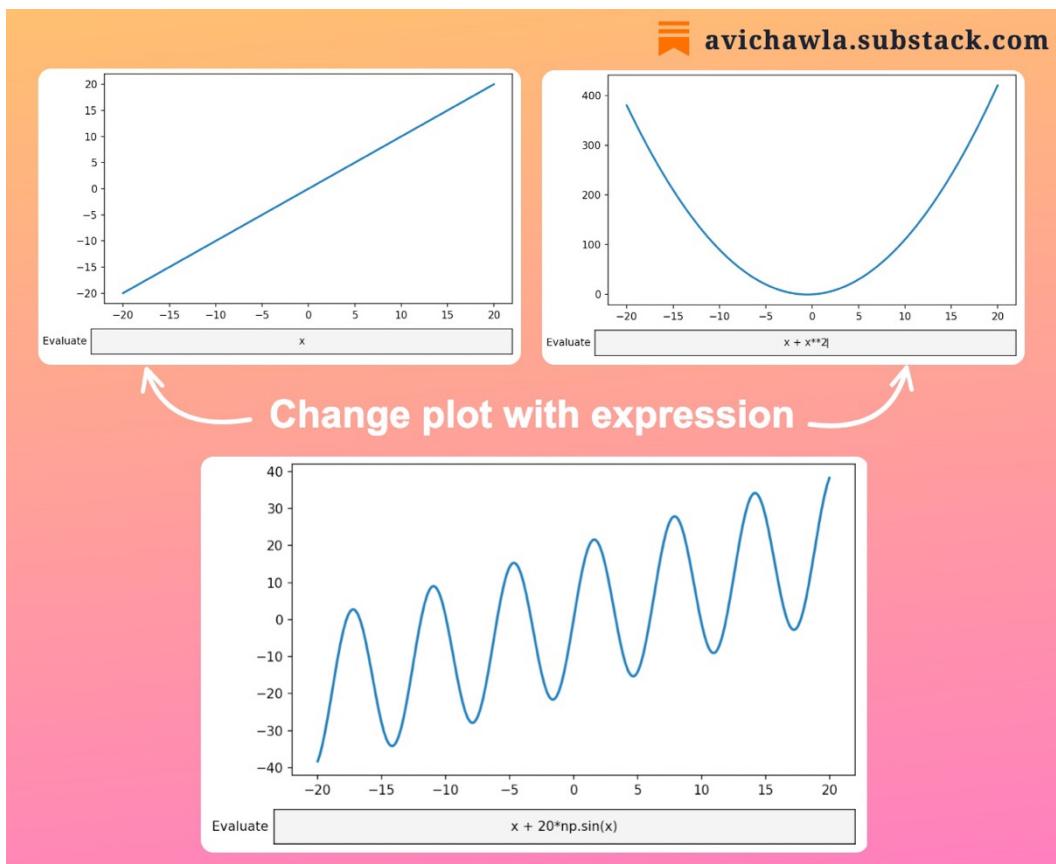
For instance, by implementing the `__new__()` method, you can apply data checks. This ensures that your program allocates memory only when certain conditions are met.



Other common use cases involve defining singleton classes (classes with only one object), creating subclasses of immutable classes such as tuples, etc.



# Who Said Matplotlib Cannot Create Interactive Plots?



👉 Please watch a video version of this post for better understanding: [Video Link](#).

In most cases, Matplotlib is used to create static plots. But very few know that it can create interactive plots too. Here's how.

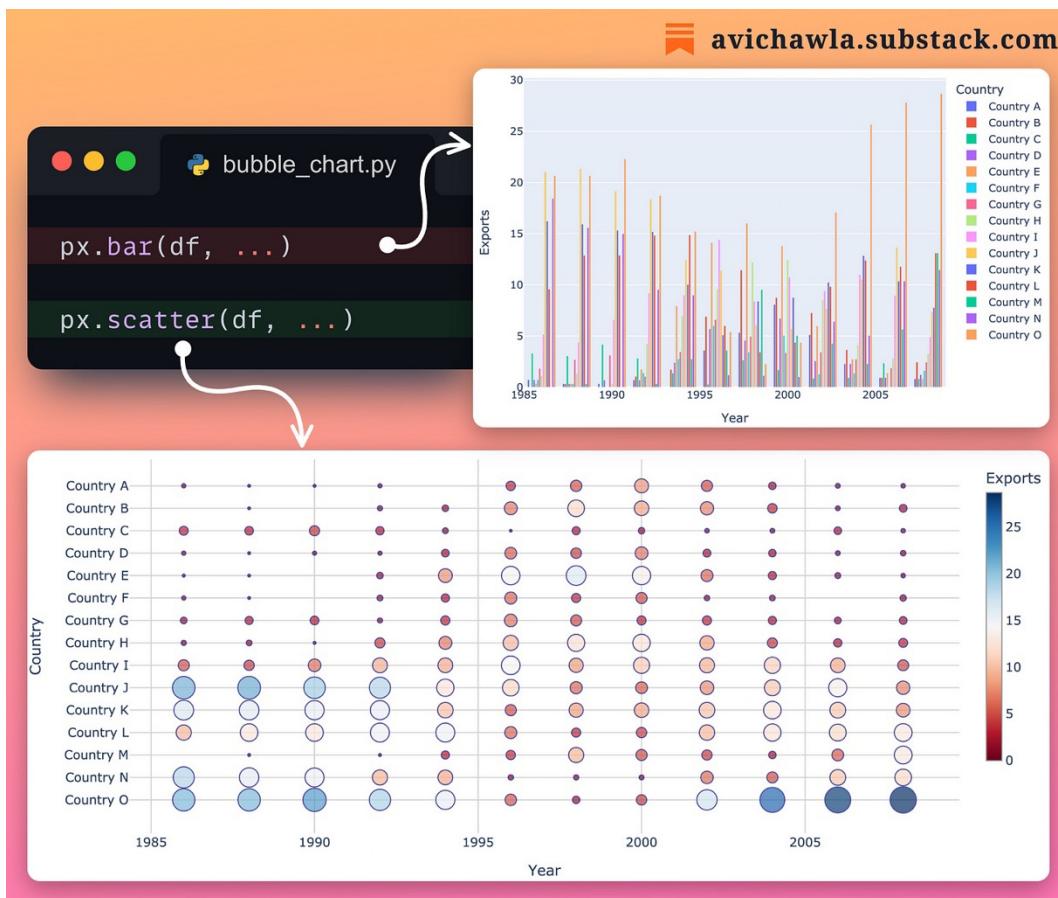
By default, Matplotlib uses the **inline** mode, which renders static plots. However, with the **%matplotlib widget** magic command, you can enable interactive backend for Matplotlib plots.

What's more, its **widgets** module offers many useful widgets. You can integrate them with your plots to make them more elegant.

Find a detailed guide here: [Matplotlib widgets](#).



# Don't Create Messy Bar Plots. Instead, Try Bubble Charts!



Bar plots often get incomprehensible and messy when we have many categories to plot.

A bubble chart can be a better choice in such cases. They are like scatter plots but with one categorical and one continuous axis.

Compared to a bar plot, they are less cluttered and offer better comprehension.

Of course, the choice of plot ultimately depends on the nature of the data and the specific insights you wish to convey.

Which plot do you typically prefer in such situations?



# You Can Add a List As a Dictionary's Key (Technically)!

```
>>> my_dict = {} ## dict
>>> my_list = [1,2,3] ## list
>>> my_dict[my_list] = True
TypeError: unhashable type: 'list'
```

unhashable  
list

```
## Inherit list class and implement __hash__ func
class MyList(list):
    def __hash__(self):
        return 0

>>> my_list = MyList([1,2,3])
>>> my_dict[my_list] = True
>>> print(my_dict)
{[1, 2, 3]: True}
```

hashable  
list

Python raises an error whenever we add a list as a dictionary's key. But do you know the technical reason behind it? Here you go.

Firstly, understand that everything in Python is an object instantiated from some class. Whenever we add an object as a dict's key, Python invokes the `__hash__` function of that object's class.

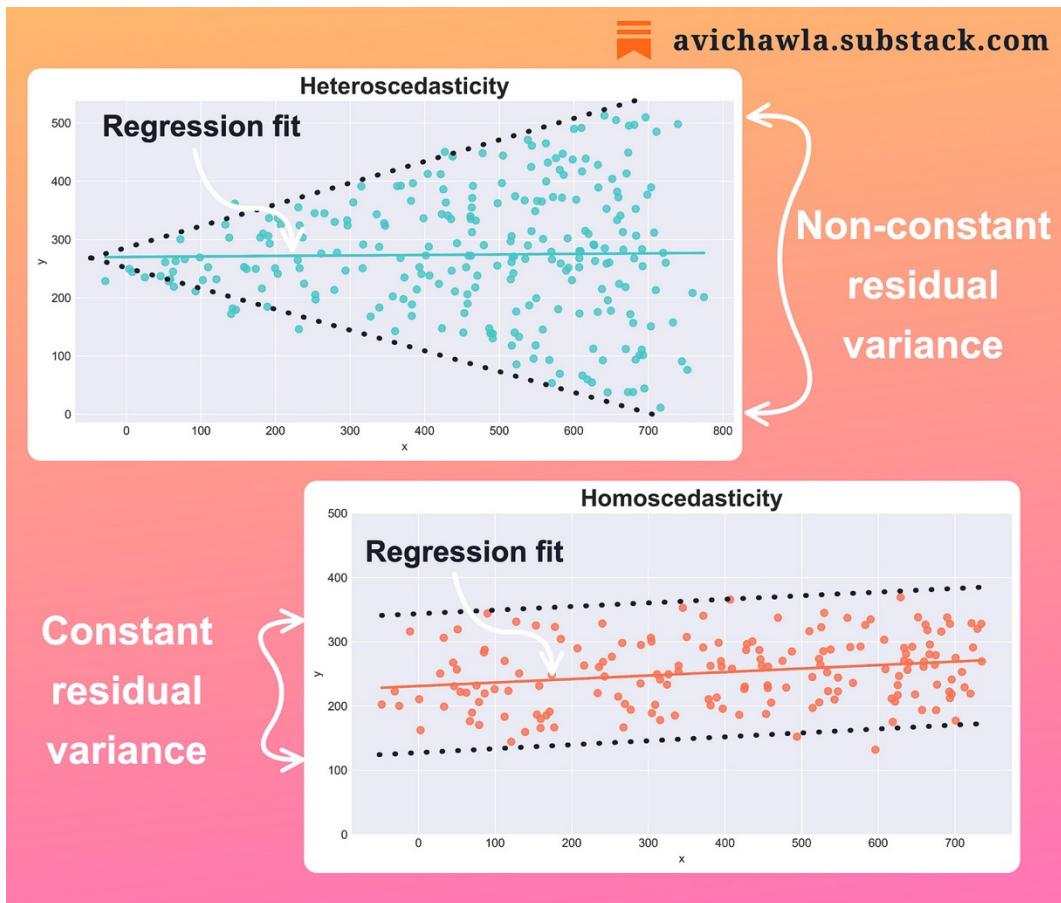
While classes of `int`, `str`, `tuple`, `frozenset`, etc. implement the `__hash__` method, it is missing from the `list` class. That is why we cannot add a list as a dictionary's key.

Thus, technically if we extend the `list` class and add this method, a list can be added as a dictionary's key.

While this makes a list hashable, it isn't recommended as it can lead to unexpected behavior in your code.



# Most ML Folks Often Neglect This While Using Linear Regression



The effectiveness of a linear regression model is determined by how well the data conforms to the algorithm's underlying assumptions.

One highly important, yet often neglected assumption of linear regression is homoscedasticity.

A dataset is homoscedastic if the variability of residuals (=actual-predicted) stays the same across the input range.

In contrast, a dataset is heteroscedastic if the residuals have non-constant variance.

Homoscedasticity is extremely critical for linear regression. This is because it ensures that our regression coefficients are reliable. Moreover, we can trust that the predictions will always stay within the same confidence interval.



## 35 Hidden Python Libraries That Are Absolute Gems



I reviewed 1,000+ Python libraries and discovered these hidden gems I never knew even existed.

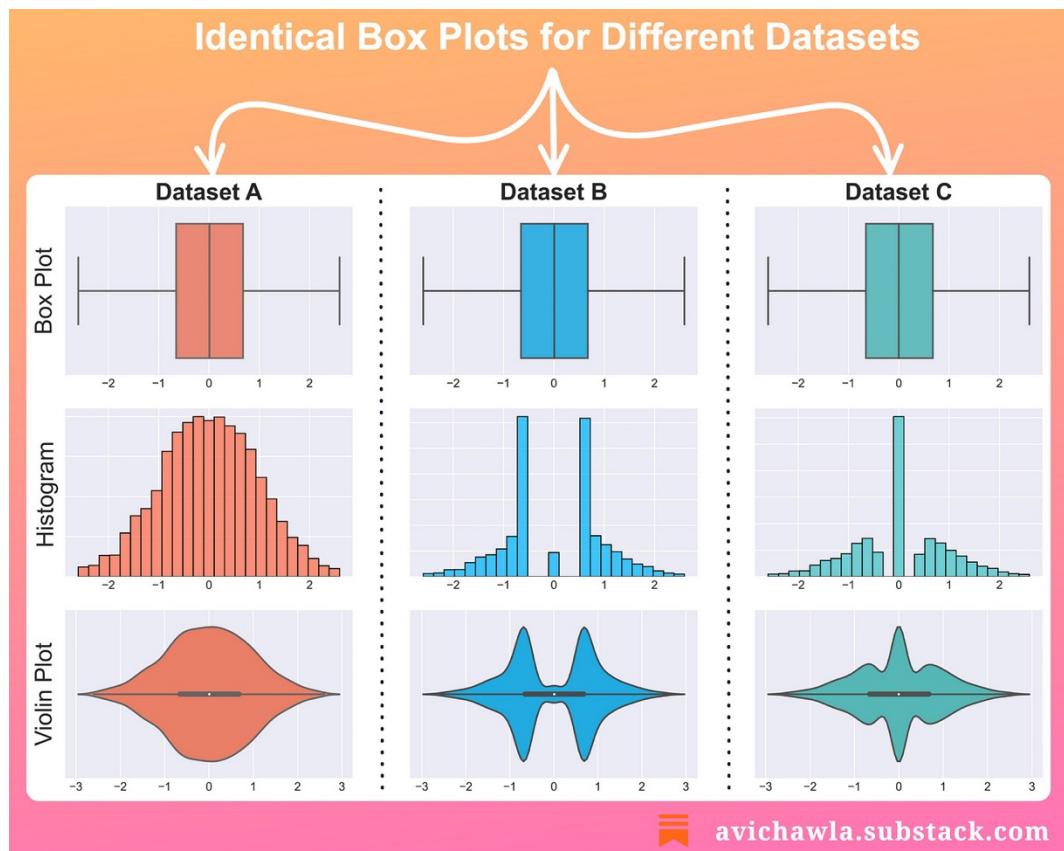
Here are some of them that will make you fall in love with Python and its versatility (even more).

Read this full list here:

<https://avichawla.substack.com/p/35-gem-py-libs>.



# Use Box Plots With Caution! They May Be Misleading.



Box plots are quite common in data analysis. But they can be misleading at times. Here's why.

A box plot is a graphical representation of just five numbers – min, first quartile, median, third quartile, and max.

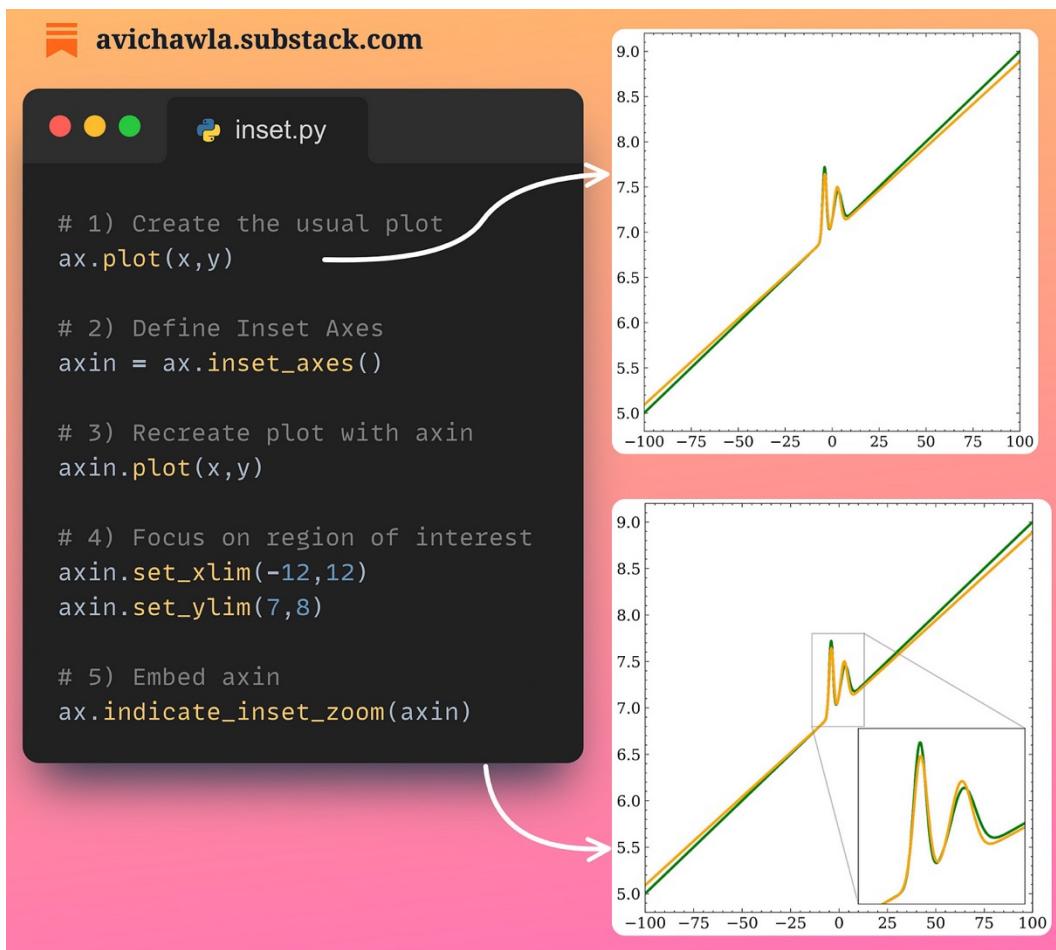
Thus, two different datasets with similar five values will produce identical box plots. This, at times, can be misleading and one may draw wrong conclusions.

The takeaway is NOT that box plots should not be used. Instead, look at the underlying distribution too. Here, histograms and violin plots can help.

Lastly, always remember that when you condense a dataset, you don't see the whole picture. You are losing essential information.



# An Underrated Technique To Create Better Data Plots



While creating visualizations, there are often certain parts that are particularly important. Yet, they may not be immediately obvious to the viewer.

A good data storyteller will always ensure that the plot guides the viewer's attention to these key areas.

One great way is to zoom in on specific regions of interest in a plot. This ensures that our plot indeed communicates what we intend it to depict.

In matplotlib, you can do so using **indicate\_inset\_zoom()**. It adds an indicator box, that can be zoomed-in for better communication.

Find more info here: [Matplotlib docs](#).



# The Pandas DataFrame Extension Every Data Scientist Has Been Waiting For

The screenshot shows a Jupyter Notebook interface with the title "Kanaries/pygwalker". A code cell contains Python code to import pandas and pygwalker, and to read a CSV file. Below the code is a visualization pane titled "Chart 1". The visualization interface includes a "Field List" sidebar with various categorical fields like date, month, season, year, holiday, hour, work yes or not, etc. On the right, there are "Filters" and "X-Axis" (set to "hour") and "Y-Axis" (set to "casual" and "registered"). The main area displays a stacked area chart with two data series: "casual" (light blue) and "registered" (light green). The x-axis represents hours from 0 to 24, and the y-axis represents counts from 0 to 60,000. The chart shows peaks around 10 AM and 6 PM. A legend indicates four seasons (fall, spring, summer, winter) and two categories for "work yes or not" (0 and 1).

**Watch a video version of this post for better understanding: [Video Link](#).**

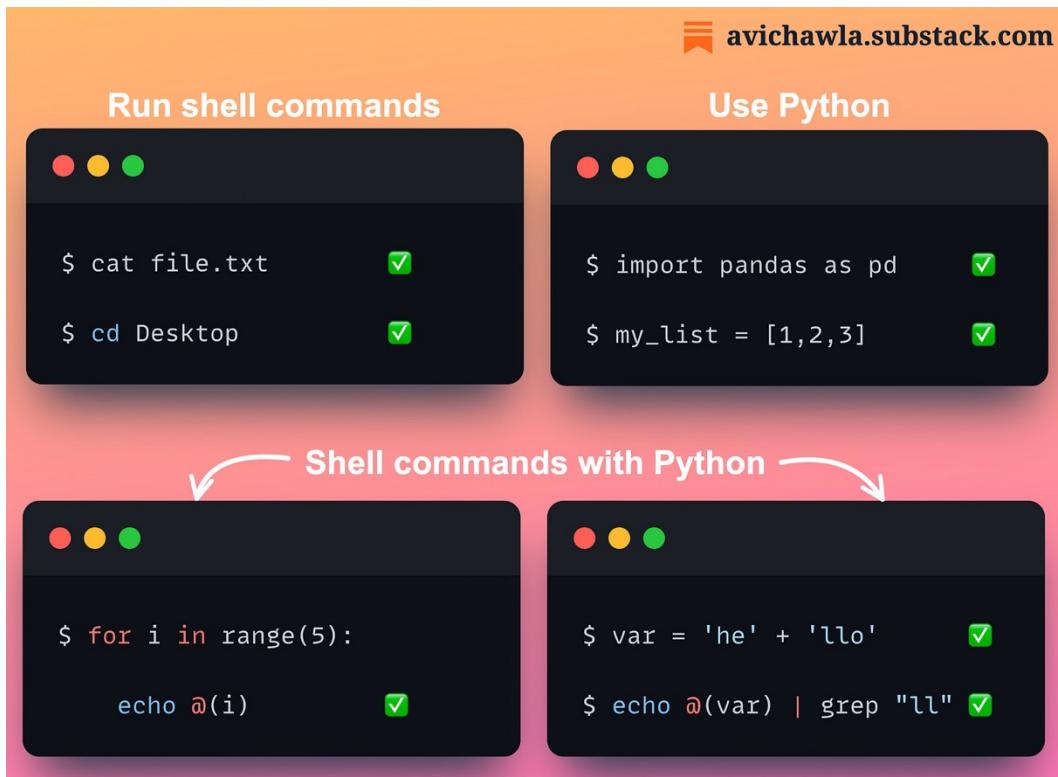
PyGWalker is an open-source alternative to Tableau that transforms pandas dataframe into a tableau-style user interface for data exploration.

It provides a tableau-like UI in Jupyter, allowing you to analyze data faster and without code.

Find more info here: [PyGWalker](#).



# Supercharge Shell With Python Using Xonsh



Traditional shells have a limitation for python users. At a time, users can either run shell commands or use IPython.

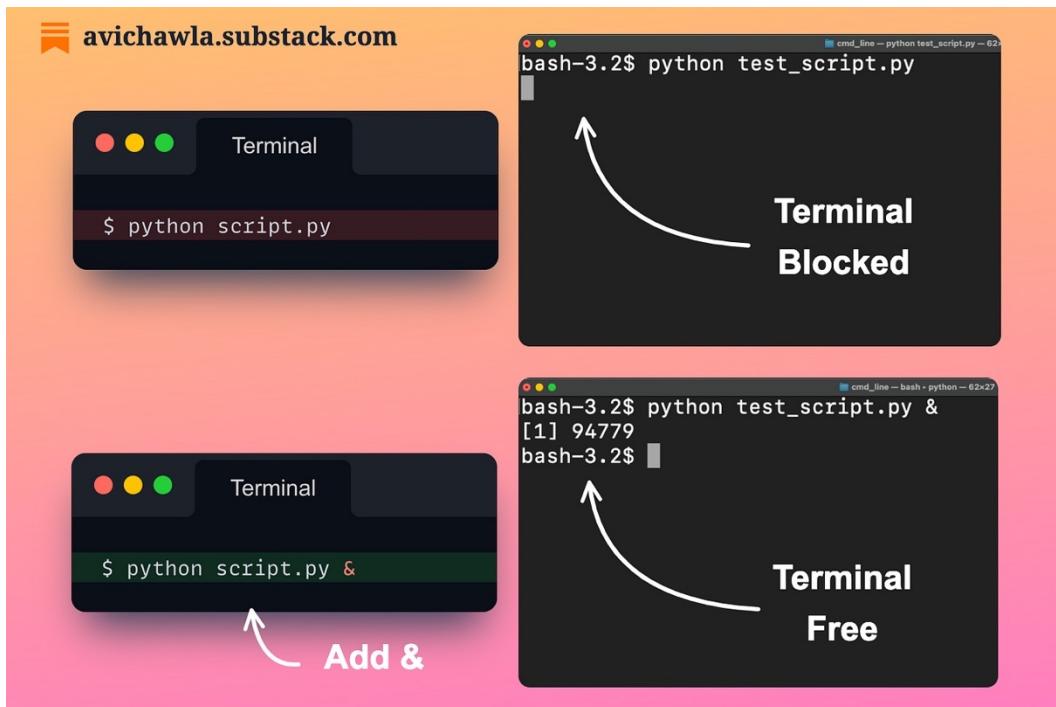
As a result, one has to open multiple terminals or switch back and forth between them in the same terminal.

Instead, try Xonsh. It combines the convenience of a traditional shell with the power of Python. Thus, you can use Python syntax as well as run shell commands in the same shell.

Find more info here: [Xonsh](#).



# Most Command-line Users Don't Know This Cool Trick About Using Terminals



**Watch a video version of this post for better understanding: [Video Link](#).**

After running a command (or script, etc.), most command-line users open a new terminal to run other commands. But that is never required.

Here's how.

When we run a program from the command line, by default, it runs in the foreground. This means you can't use the terminal until the program has been completed.

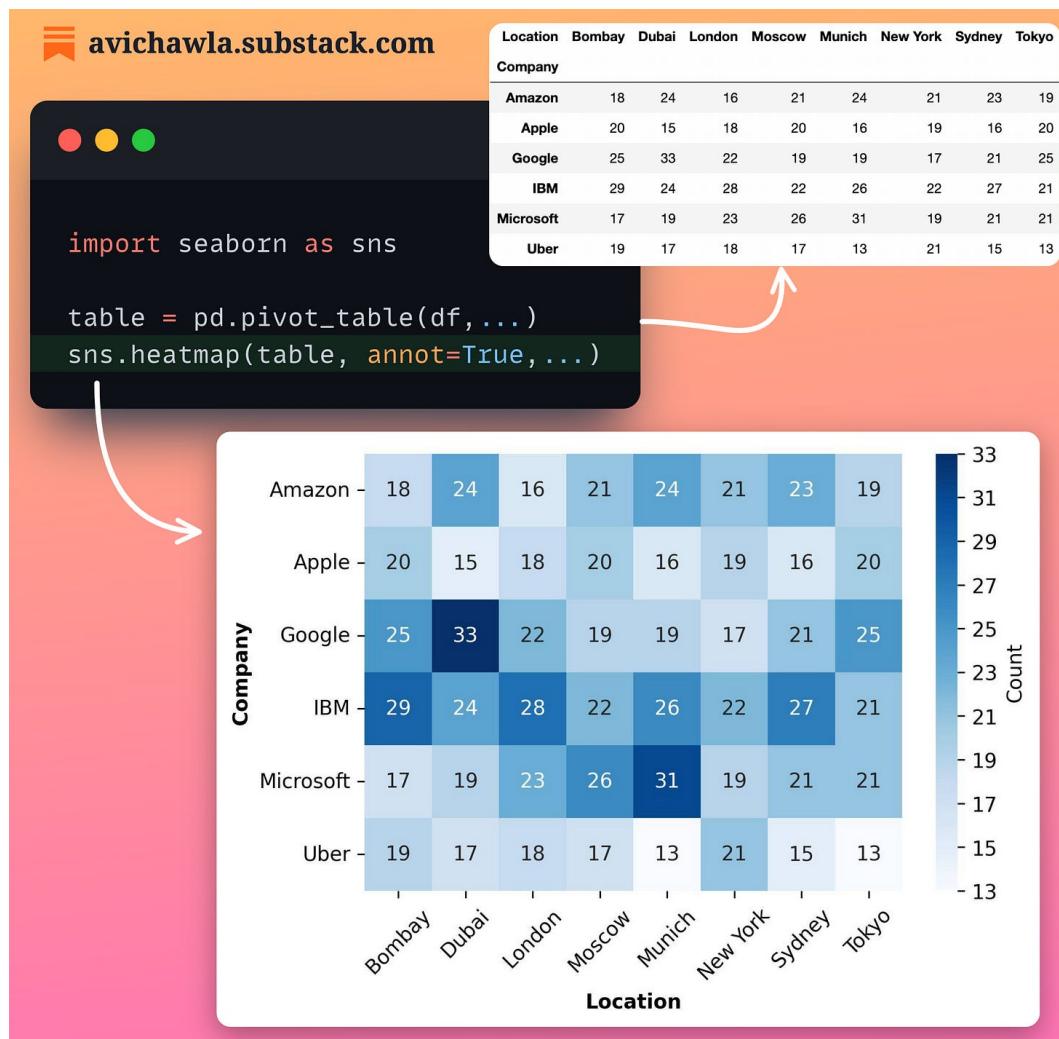
However, if you add '**&**' at the end of the command, the program will run in the background and instantly free the terminal.

This way, you can use the same terminal to run another command.

To bring the program back to the foreground, use the '**fg**' command.



# A Simple Trick to Make The Most Out of Pivot Tables in Pandas



Pivot tables are pretty common for data exploration. Yet, analyzing raw figures is tedious and challenging. What's more, one may miss out on some crucial insights about the data.

Instead, enrich your pivot tables with heatmaps. The color encodings make it easier to analyze the data and determine patterns.



# Why Python Does Not Offer True OOP Encapsulation

The screenshot shows a Jupyter Notebook interface with two code cells. The top cell contains the definition of a class `MyClass` with three attributes: `public_attr`, `_protected_attr`, and `__private_attr`. The bottom cell shows the creation of an instance `my_obj` and the access of these attributes via the dot notation. Three arrows on the right side point from the text labels to the corresponding code examples.

```
class MyClass:  
    def __init__(self):  
        self.public_attr = "I'm public"      # 0 underscores  
        self._protected_attr = "I'm protected" # 1 underscore  
        self.__private_attr = "I'm private"   # 2 underscores  
  
my_obj = MyClass()  
  
=> my_obj.public_attr  
"I'm public"  
  
=> my_obj._protected_attr  
"I'm protected"  
  
=> my_obj._MyClass__private_attr  
"I'm private"
```

**Public member accessible**

**Protected member accessible**

**Private member accessible with name mangling**

Using access modifiers (public, protected, and private) is fundamental to encapsulation in OOP. Yet, Python, in some way, fails to deliver true encapsulation.

By definition, a public member is accessible everywhere. A private member can only be accessed inside the base class. A protected member is accessible inside the base class and child class(es).

But, with Python, there are no such strict enforcements.

Thus, protected members behave exactly like public members. What's more, private members can be accessed outside the class using name mangling.

As a programmer, remember that encapsulation in Python mainly relies on conventions. Thus, it is the responsibility of the programmer to follow them.



# Never Worry About Parsing Errors Again While Reading CSV with Pandas

```
In [1]: !cat file.csv
Name,Amount
Alice,$300
Bob,$1\,000
Charlie,$200
```

Separator appears  
in value

```
In [2]: pd.read_csv("file.csv")
## ParserError: Error tokenizing data. C error:
## Expected 2 fields in line 3, saw 3
```

```
In [3]: import clevercsv
clevercsv.read_dataframe("file.csv")
```

Out[3]:

	Name	Amount
0	Alice	\$300
1	Bob	\$1,000
2	Charlie	\$200

 avichawla.substack.com

Pandas isn't smart (yet) to read messy CSV files.

Its `read_csv` method assumes the data source to be in a standard tabular format. Thus, any irregularity in data raises parsing errors, which may require manual intervention.

Instead, try CleverCSV. It detects the format of CSVs and makes it easier to load them, saving you tons of time.

Find more info here: [CleverCSV](#).



# An Interesting and Lesser-Known Way To Create Plots Using Pandas

```
from base64 import b64encode
from io import BytesIO

def create_hist(data):
    fig, ax = plt.subplots(figsize=(2, 0.5))
    ax.hist(data, bins=10)
    ax.axis('off')
    plt.close(fig)

    img = BytesIO() # create Bytes Object
    fig.savefig(img) # Save Image to Bytes Object
    encoded = b64encode(img.getvalue()) # Encode object as base64 byte string
    decoded = encoded.decode('utf-8') # Decode to utf-8
    return f'' # Return HTML tag
```

```
df['Last 7 Days'] = df['Price History'].apply(create_line)
df['Trade Volume'] = df['Price History'].apply(create_hist)
```

```
HTML(df.to_html(escape=False))
```



Whenever you print/display a DataFrame in Jupyter, it is rendered using HTML and CSS. This allows us to format the output just like any other web page.

One interesting way is to embed inline plots which appear as a column of a dataframe.

In the above snippet, we first create a plot as we usually do. Next, we return the <img> HTML tag with its source as the plot. Lastly, we render the dataframe as HTML.

Find the code for this tip here: [Notebook](#).



# Most Python Programmers Don't Know This About Python For-loops



Often when we use a for-loop in Python, we tend not to modify the loop variable inside the loop.

The impulse typically comes from acquaintance with other programming languages like C++ and Java.

But for-loops don't work that way in Python. Modifying the loop variable has no effect on the iteration.

This is because, before every iteration, Python unpacks the next item provided by iterable (`range(5)`) and assigns it to the loop variable (`num`).

Thus, any changes to the loop variable are replaced by the new value coming from the iterable.



# How To Enable Function Overloading In Python

python interpreter only considers the latest definition of add() function

```
def add(x:int, y:int):
    return x + y

def add(x:int, y:int, z:int):
    return x + y + z

>>> add(1,2)
TypeError: add() missing 1 required positional argument: 'z'
```

from multipledispatch import dispatch

```
@dispatch(int, int)
def add(x, y):
    return x + y

@dispatch(int, int, int)
def add(x, y, z):
    return x + y + z

>>> add(1,2)           >>> add(1,2,3)
3                         6
```

dispatch decorator enables function overloading

Python has no native support for function overloading. Yet, there's a quick solution to it.

Function overloading (having multiple functions with the same name but different number/type of parameters) is one of the core ideas behind polymorphism in OOP.

But if you have many functions with the same name, python only considers the latest definition. This restricts writing polymorphic code.

Despite this limitation, the `dispatch` decorator allows you to leverage function overloading.

Find more info here: [Multipledispatch](#).



# Generate Helpful Hints As You Write Your Pandas Code

The screenshot shows a Jupyter Notebook interface with three code cells and their corresponding Dovpanda hints:

- In [4]:** `import dovpanda`
- In [5]:** `iter_df = df.iterrows()`  
A hint box appears: `df.iterrows` is not recommended. Essentially it is very similar to iterating the rows of the frames in a loop. In the majority of cases, there are better alternatives that utilize pandas' vector operation  
Line 1: `iter_df = df.iterrows()`
- In [6]:** `df["new_col"] = df.apply(apply_func)`  
A hint box appears: `df.apply` is not recommended. Essentially it is very similar to iterating the rows of the frames in a loop. In the majority of cases, there are better alternatives that utilize pandas' vector operation  
Line 1: `df["new_col"] = df.apply(apply_func)`
- In [7]:** `merged_df = pd.concat((df, df))`  
A hint box appears: All dataframes have the same columns and same number of rows. Pay attention, your axis is 0 which concatenates vertically  
Line 1: `merged_df = pd.concat((df, df))`
- In [8]:** `merged_df = pd.concat((df, df), ignore_index=True)`  
A hint box appears: After concatenation you have duplicated indices - pay attention  
Line 1: `merged_df = pd.concat((df, df), ignore_index=True)`

When manipulating a dataframe, at times, one may be using unoptimized methods. What's more, errors introduced into the data can easily go unnoticed.

To get hints and directions about your data/code, try Dovpanda. It works as a companion for your Pandas code. As a result, it gives suggestions/warnings about your data manipulation steps.

P.S. When you will import Dovpanda, you will likely get an error. Ignore it and proceed with using Pandas. You will still receive suggestions from Dovpanda.

Find more info here: [Dovpandas](#).



# Speedup NumPy Methods 25x With Bottleneck

The screenshot shows a Mac desktop with two terminal windows side-by-side. The top window contains Python code to import bottleneck and numpy, and generate a random array:

```
import bottleneck as bn
import numpy as np

arr = np.random.random((1000,))
```

The bottom window has two tabs: "numpy.py" and "bottleneck.py". Both tabs show command-line sessions. The "numpy.py" tab lists standard NumPy methods with their run-times:

- >>> np.sum(arr)  
## Run-time: 870 μs
- >>> np.mean(arr)  
## Run-time: 477 μs
- >>> np.std(arr)  
## Run-time: 687 μs
- >>> np.median(arr)  
## Run-time: 1.58 ms
- >>> np.max(arr)  
## Run-time: 1.26 ms

The "bottleneck.py" tab shows the corresponding bottleneck methods, each with a green background indicating they are faster:

- >>> bn.nansum(arr)  
## Run-time: 33.9 μs (25x Faster)
- >>> bn.nanmean(arr)  
## Run-time: 21 μs (22x Faster)
- >>> bn.nanstd(arr)  
## Run-time: 175 μs (4x Faster)
- >>> bn.nanmedian(arr)  
## Run-time: 0.43 ms (4x Faster)
- >>> bn.nanmax(arr)  
## Run-time: 0.46 ms (3x Faster)

Arrows from the "numpy.py" rows point to the corresponding "bottleneck.py" rows.

NumPy's methods are already highly optimized for performance. Yet, here's how you can further speed them up.

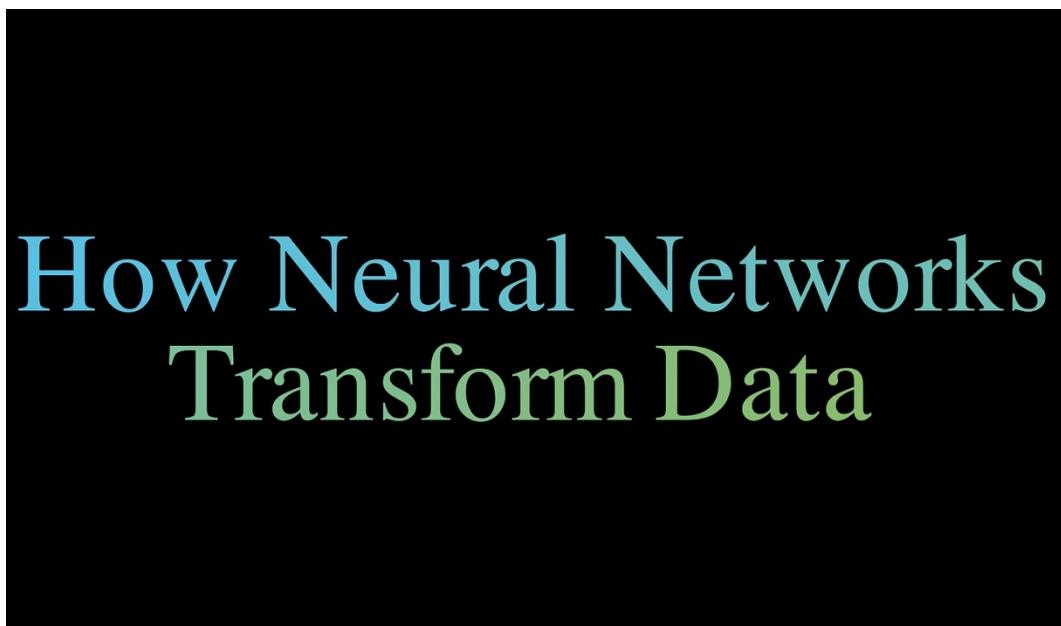
Bottleneck provides a suite of optimized implementations of NumPy methods.

Bottleneck is especially efficient for arrays with NaN values where performance boost can reach up to 100-120x.

Find more info here: [Bottleneck](#).



## Visualizing The Data Transformation of a Neural Network



If you struggle to comprehend how a neural network learns complex non-linear data, I have created an animation that will surely help.

Please find the video here: [\*\*Neural Network Animation.\*\*](#)

For linearly inseparable data, the task boils down to projecting the data to a space where it becomes linearly separable.

Now, either you could do this manually by adding relevant features that will transform your data to a linear separable form. Consider concentric circles for instance. Passing a square of  $(x,y)$  coordinates as a feature will do this job.

But in most cases, the transformation is unknown or complex to figure out. Thus, non-linear activation functions are considered the best bet, and a neural network is allowed to figure out this "non-linear to linear transformation" on its own.

As shown in the animation, if we tweak the neural network by adding a 2D layer right before the output, and visualize this transformation, we see that the neural network has learned to linearly separate the data. We add a layer 2D because it is easy to visualize.

This linearly separable data can be easily classified by the last layer. To put it another way, the last layer is analogous to a logistic regression model which is given a linear separable input.

The code for this visualization experiment is available here: [GitHub](#).



# Never Refactor Your Code Manually Again. Instead, Use Sourcery!

Before Refactoring

```
def is_special_number(number):
    if number == 7:
        return True
    elif number == 18:
        return True
    else:
        return False
```

Command Line

```
$ sourcery review --in-place my_code.py
```

After Refactoring

```
def is_special_number(number):
    return number in [7, 18]
```

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Refactoring code is an important step in pipeline development. Yet, manual refactoring takes additional time for testing as one might unknowingly introduce errors.

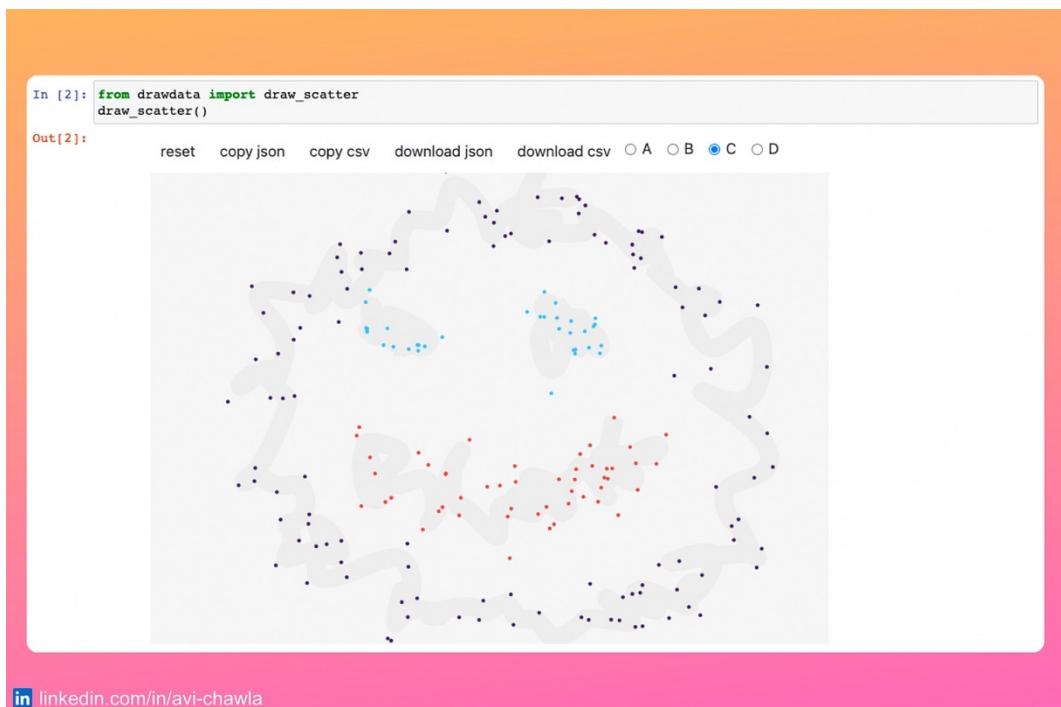
Instead, use Sourcery. It's an automated refactoring tool that makes your code elegant, concise, and Pythonic in no time.

With Sourcery, you can refactor code from the command line, as an IDE plugin in VS Code and PyCharm, pre-commit, etc.

Find more info here: [Sourcery](#).



# Draw The Data You Are Looking For In Seconds



Please watch a video version of this post for better understanding: [Video Link](#).

Often when you want data of some specific shape, programmatically generating it can be a tedious and time-consuming task.

Instead, use `drawdata`. This allows you to draw any 2D dataset in a notebook and export it. Besides a scatter plot, it can also create histogram and line plot

Find more info here: [Drawdata](#).



# Style Matplotlib Plots To Make Them More Attractive



Matplotlib offers close to 50 different styles to customize the plot's appearance.

To alter the plot's style, select a style from **plt.style.available** and create the plot as you originally would.

Find more info about styling here: [Docs](#).



# Speed-up Parquet I/O of Pandas by 5x

The image shows a Mac desktop with two terminal windows side-by-side. The background is a gradient from orange to pink.

**Terminal 1 (Left):**

- Icon: Python (pandas.py)
- Title: file.parquet:
- Text:

```
import pandas as pd

df = pd.read_parquet("file.parquet")
# Run-time: 41s
```

**Terminal 2 (Right):**

- Icon: Python (fastparquet.py)
- Title: file.parquet:
- Text:

```
from fastparquet import ParquetFile

pf = ParquetFile('file.parquet')
df = pf.to_pandas()
# Run-time: 8.1s
```

A white arrow points from the text "5x Faster" to the opening brace of the fastparquet code block.

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Dataframes are often stored in parquet files and read using Pandas' **read\_parquet()** method.

Rather than using Pandas, which relies on a single-core, use fastparquet. It offers immense speedups for I/O on parquet files using parallel processing.

Find more info here: [Docs](#).



## 40 Open-Source Tools to Supercharge Your Pandas Workflow



Pandas receives over [3M downloads per day](#). But 99% of its users are not using it to its full potential.

I discovered these open-source gems that will immensely supercharge your Pandas workflow the moment you start using them.

Read this list here: <https://avichawla.substack.com/p/37-open-source-tools-to-supercharge-pandas>.



# Stop Using The Describe Method in Pandas. Instead, use Skimpy.

```
from skimpy import skim
skim(df)
```

Data Summary		Data Types		skimpy summary						
dataframe	Values	Column Type	Count	Categories						
Number of rows	1000	float64	3							
Number of columns	10	category	2							
		datetime64	2							
		int64	1							
		bool	1							
		string	1							
column_name	NA	NA %	mean	sd	p0	p25	p75	p100	hist	
length	0	0	0.5	0.36	1.6e-06	0.13	0.86	1		
width	0	0	2	1.9	0.0021	0.6	3	14		
depth	0	0	10	3.2	2	8	12	20		
rnd	120	12	-0.02	1	-2.8	-0.74	0.66	3.7		
column_name	NA	NA %	ordered	unique						
class	0	0	False	2						
location	1	0.1	False	5						
column_name	NA	NA %	first	last	frequency					
date	0	0	2018-01-31	2101-04-30	M					
date_no_freq	3	0.3	1992-01-05	2023-03-04	None					
column_name	NA	NA %	words per row	total words						
text	6	0.6		5.8					5800	
column_name	true	true rate		hist						
booly_col		520		0.52						

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Supercharge the describe method in Pandas.

Skimpy is a lightweight tool for summarizing Pandas dataframes. In a single line of code, it generates a richer statistical summary than the `describe()` method.

What's more, the summary is grouped by datatypes for efficient analysis. You can use Skimpy from the command line too.

Find more info here: [Docs](#).



# The Right Way to Roll Out Library Updates in Python

```
my_library.py
from deprecated import deprecated
@deprecated(reason="old_function will be \
    deprecated in the next \
    release. Use new_function.")
def old_function():
    ...

project.py
old_value = old_function()
DeprecationWarning: Call to deprecated function
old_function. (old_function will be deprecated
in the next release. Use new_function.)
```

my\_library.py

project.py

Add decorator

Prints warning

linkedin.com/in/avi-chawla

While developing a library, authors may decide to remove some functions/methods/classes. But instantly rolling the update without any prior warning isn't a good practice.

This is because many users may still be using the old methods and they may need time to update their code.

Using the **deprecated** decorator, one can convey a warning to the users about the update. This allows them to update their code before it becomes outdated.

Find more info here: [GitHub](#).



# Simple One-Liners to Preview a Decision Tree Using Sklearn

```
my_tree = DecisionTreeClassifier()
my_tree.fit(X, y)

from sklearn.tree import plot_tree, export_text
```

```
plot_tree(my_tree, feature_names=features,
          class_names=classes, filled=True)
```

**Method 1**

```
graph TD
    Root["petal_width <= 0.8  
gini = 0.667  
samples = 150  
value = [50, 50, 50]  
class = setosa"]
    Root --> L1L["gini = 0.0  
samples = 50  
value = [50, 0, 0]  
class = setosa"]
    Root --> L1R["petal_width <= 1.75  
gini = 0.5  
samples = 100  
value = [0, 50, 50]  
class = versicolor"]
    L1L --> L2L["gini = 0.168  
samples = 54  
value = [0, 49, 5]  
class = versicolor"]
    L1R --> L2R["gini = 0.043  
samples = 46  
value = [0, 1, 45]  
class = virginica"]
```

```
print(export_text(my_tree, feature_names=features))
```

**Method 2**

```
--- petal_width <= 0.80
    --- class: setosa
--- petal_width >  0.80
    |--- petal_width <= 1.75
        |--- class: versicolor
        |--- petal_width > 1.75
            |--- class: virginica
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

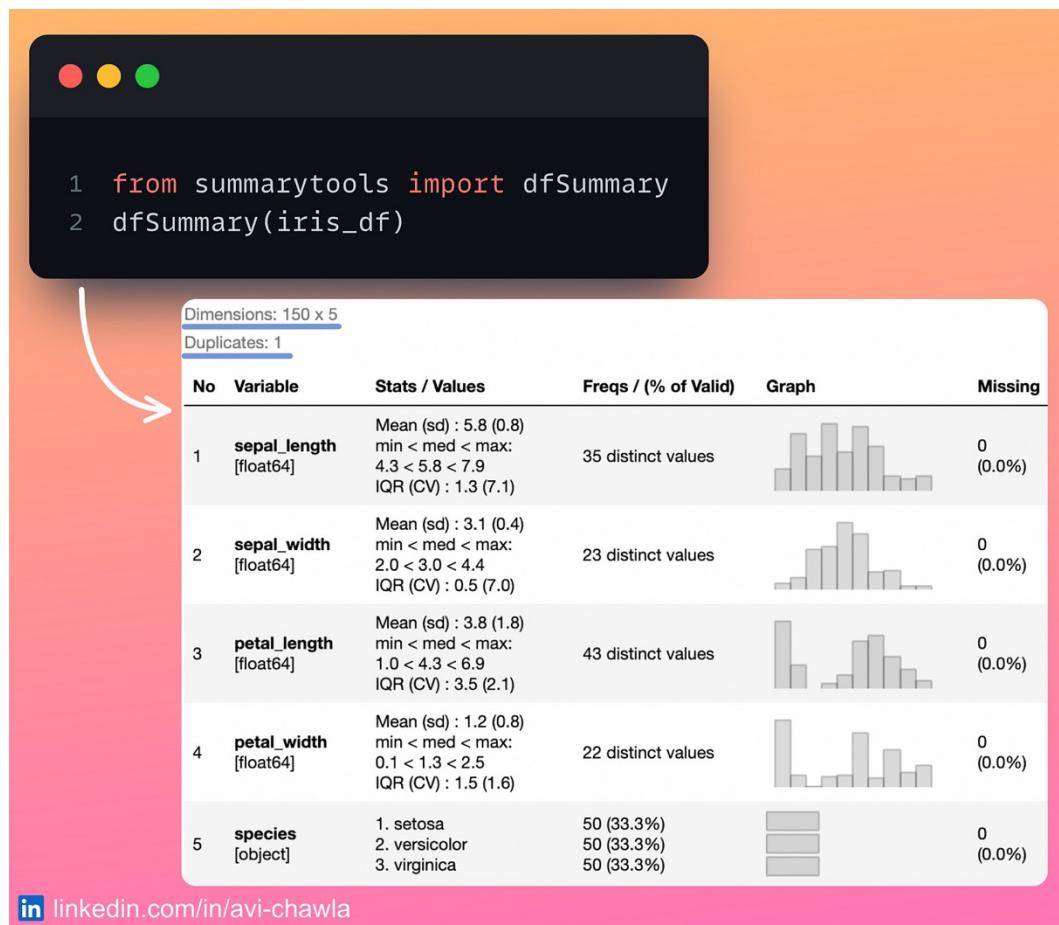
If you want to preview a decision tree, sklearn provides two simple methods to do so.

1. [plot\\_tree](#) creates a graphical representation of a decision tree.
2. [export\\_text](#) builds a text report showing the rules of a decision tree.

This is typically used to understand the rules learned by a decision tree and gaining a better understanding of the behavior of a decision tree model.



# Stop Using The Describe Method in Pandas. Instead, use Summarytools.



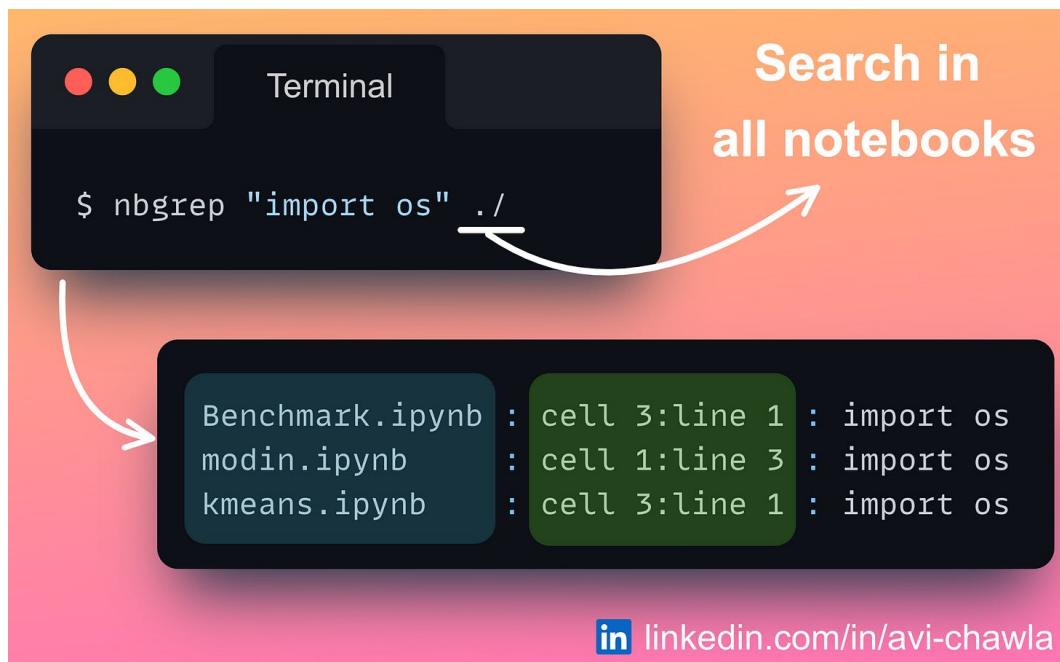
Summarytools is a simple EDA tool that gives a richer summary than **describe()** method. In a single line of code, it generates a standardized and comprehensive data summary.

The summary includes column statistics, frequency, distribution chart, and missing stats.

Find more info here: [Summary Tools](#).



# Never Search Jupyter Notebooks Manually Again To Find Your Code



Have you ever struggled to recall the specific Jupyter notebook in which you wrote some code? Here's a quick trick to save plenty of manual work and time.

**nbcommands** provides a bunch of commands to interact with Jupyter from the terminal.

For instance, you can search for code, preview a few cells, merge notebooks, and many more.

Find more info here: [GitHub](#).



# F-strings Are Much More Versatile Than You Think

The diagram illustrates two sections of f-string usage:

**Formatting:**

```
>>> print(f"3 decimals: {number:.3f}")
>>> print(f"5 digits: {number:05}")
>>> print(f"scientific: {number:e}")
```

Output:

```
3 decimals: 205.000
5 digits: 00205
scientific: 2.050000e+02
```

**Converting:**

```
>>> print(f"binary: {number:b}")
>>> print(f"hex: {number:#0x}")
>>> print(f"octal: {number:o}")
```

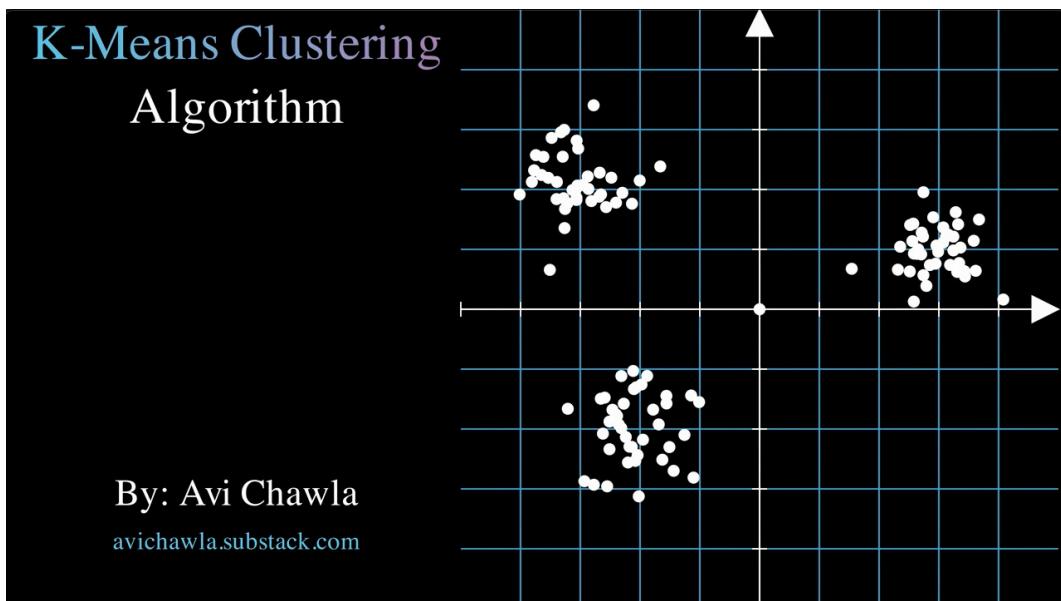
Output:

```
binary: 11001101
hex: 0xcd
octal: 315
```

Here are 6 lesser-known ways to format/convert a number using f-strings. What is your favorite f-string hack?



# Is This The Best Animated Guide To KMeans Ever?



Have you ever struggled with understanding KMeans? How it works, how are the data points assigned to a centroid, or how do the centroids move?

If yes, let me help.

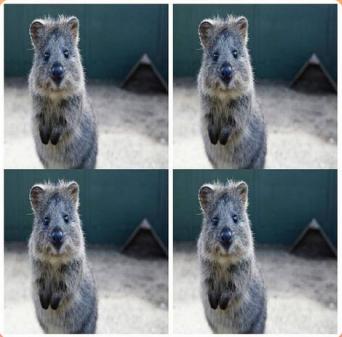
I created a beautiful animation using Manim to help you build an intuitive understanding of the algorithm.

Please find this video here: [\*\*Video Link.\*\*](#)



# An Effective Yet Underrated Technique To Improve Model Performance

**Original Images**



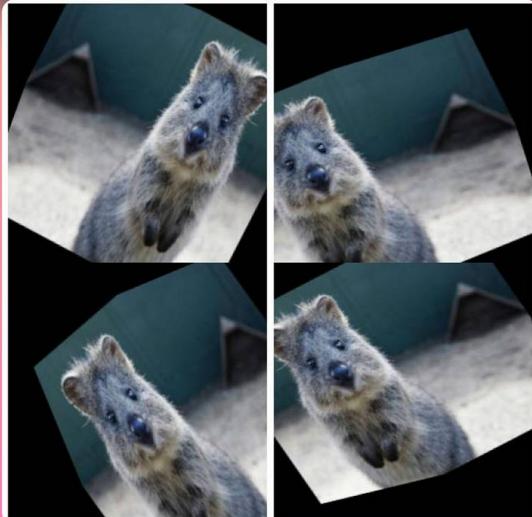
A row of four original images of a quokka, a small, dark-furred marsupial.

```
import imgaug.augmenters as iaa

seq = iaa.Sequential([
    iaa.Fliplr(0.5), # horizontal flip
    iaa.Rotate((-40,40)), # Rotate
    ...
])

images_aug = seq(images=images)
```

**Augmented Images**



A row of four augmented images of a quokka, showing horizontal flips and rotations applied to the original images.

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Robust ML models are driven by diverse training data. Here's a simple yet highly effective technique that can help you create a diverse dataset and increase model performance.

One way to increase data diversity is using data augmentation.

The idea is to create new samples by transforming the available samples. This can prevent overfitting, improve performance, and build robust models.

For images, you can use imgaug (linked in comments). It provides a variety of augmentation techniques such as flipping, rotating, scaling, adding noise to images, and many more.

Find more info: [ImgAug](#).



# Create Data Plots Right From The Terminal

```
>>> from bashplotlib.histogram import plot_hist
>>> np_arr = np.random.normal(size=1000)
>>> plot_hist(np_arr, bincount=50)

54|          o
51|      oo  oo
48|      ooo  ooo  o
45|      ooo  ooo  o
43|      oooo  ooo  o
40|      ooooooo  oo
37|      oo  ooooooo  oooooo
34|      oo  ooooooo  oooooo
31|      ooooooo  oooooo
29|      ooooooo  oooooo
26|      ooooooo  oooooo
23|      ooooooo  oooooo
20|      ooooooo  oooooo
17|      ooooooo  oooooo  o
15|      ooooooo  oooooo  o
12|      ooooooo  oooooo  o
9|      ooooooo  oooooo
6|      oo  ooooooo  oooooo
3|  o   o  ooooooo  oooooo
1| o  o  ooooooo  oooooo  o
```

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

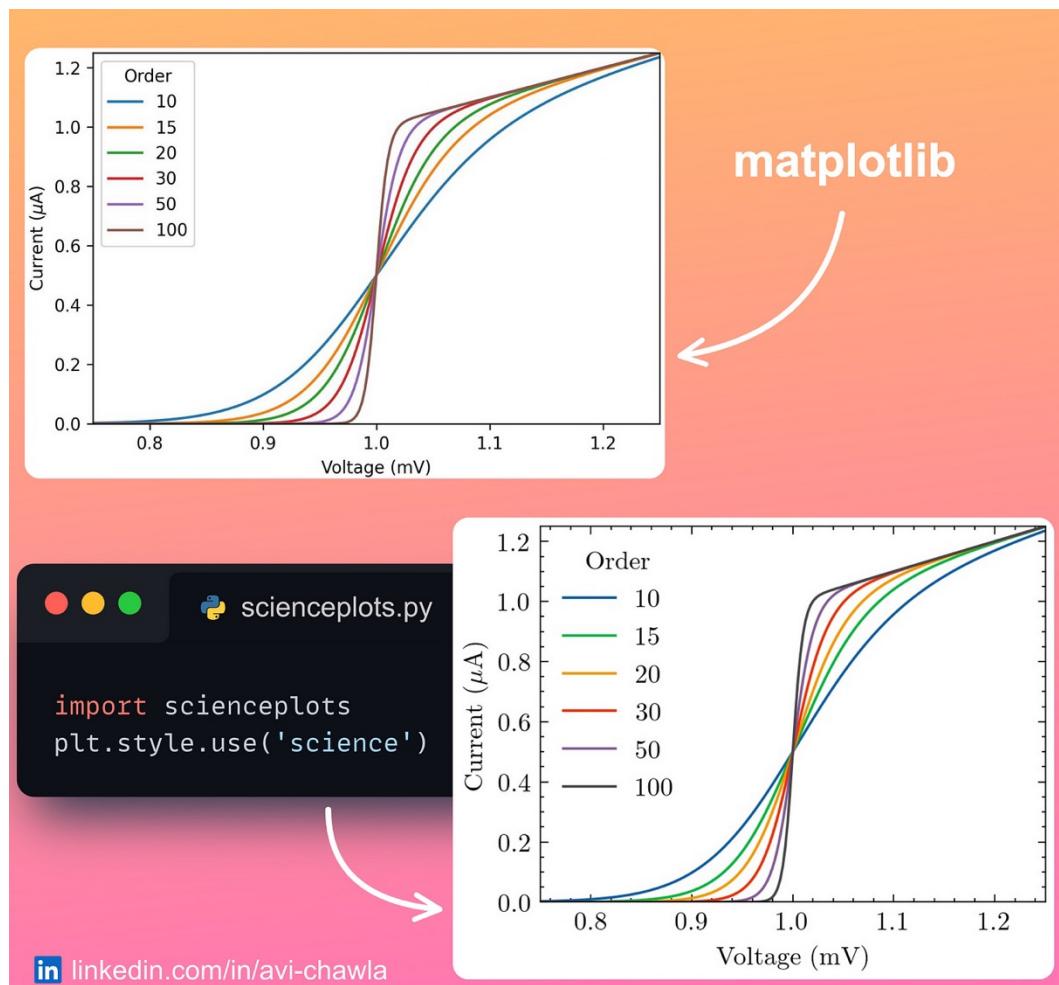
Visualizing data can get tough when you don't have access to a GUI. But here's what can help.

Bashplotlib offers a quick and easy way to make basic plots right from the terminal. Being pure python, you can quickly install it anywhere using pip and visualize your data.

Find more info here: [Bashplotlib](#).



# Make Your Matplotlib Plots More Professional



The default matplotlib plots are pretty basic in style and thus, may not be the apt choice always. Here's how you can make them appealing.

To create professional-looking and attractive plots for presentations, reports, or scientific papers, try Science Plots.

Adding just two lines of code completely transforms the plot's appearance.

Find more info here: [GitHub](#).



## 37 Hidden Python Libraries That Are Absolute Gems



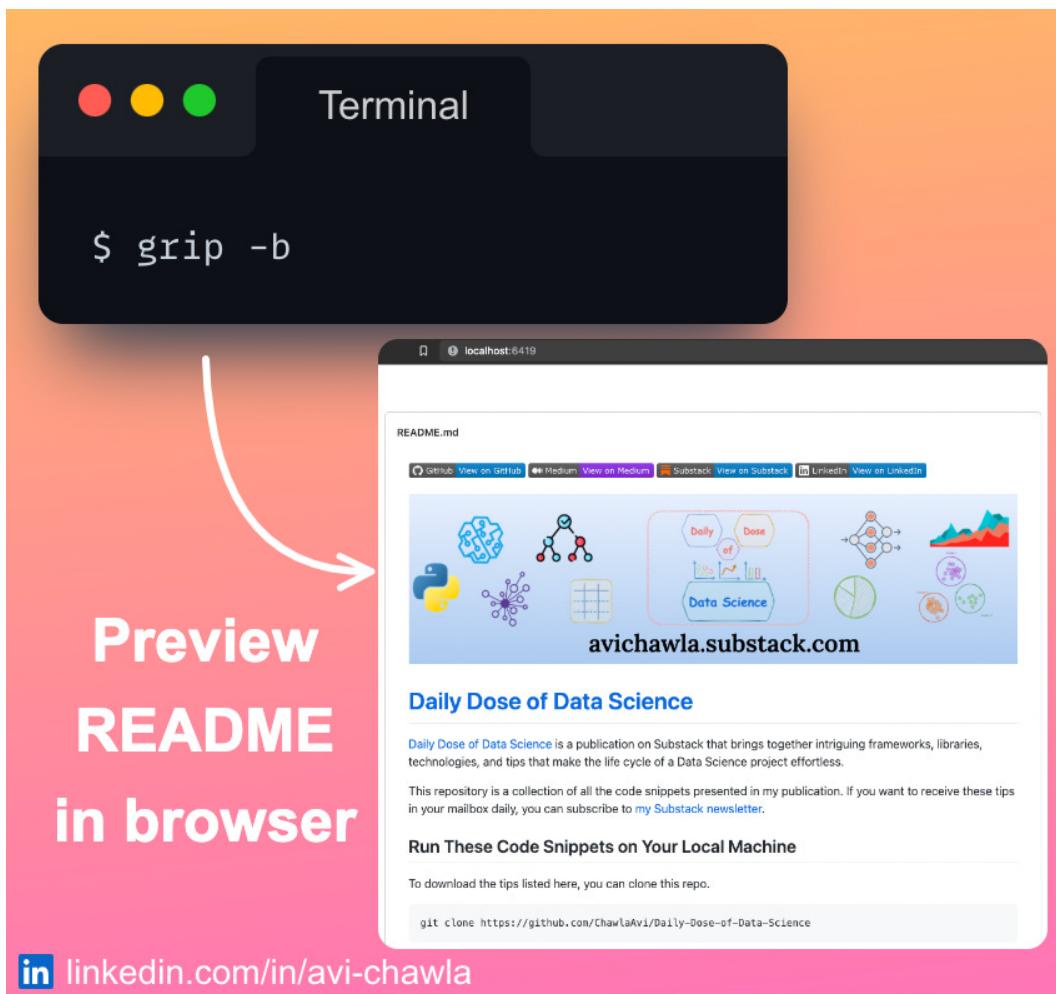
I reviewed 1,000+ Python libraries and discovered these hidden gems I never knew even existed.

Here are some of them that will make you fall in love with Python' and its versatility (even more).

Read this list here: <https://avichawla.substack.com/p/gem-libraries>.



# Preview Your README File Locally In GitHub Style



Please watch a video version for better understanding: [Video Link](#).

Have you ever wanted to preview a README file before committing it to GitHub? Here's how to do it.

Grip is a command-line tool that allows you to render a README file as it will appear on GitHub. This is extremely useful as sometimes one may want to preview the file before pushing it to GitHub.

What's more, editing the README instantly reflects in the browser without any page refresh.

Read more: [Grip](#).



# Pandas and NumPy Return Different Values for Standard Deviation. Why?

```
import numpy as np
import pandas as pd

X = np.arange(20)
df = pd.DataFrame(X)

print(f"NumPy : {np.std(X)}")
print(f"Pandas: {df.std()}")
```

**std-dev using  
Pandas and  
NumPy**

**different  
output**

NumPy : 5.766  
Pandas: 5.916

[in](https://linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Pandas assumes that the data is a sample of the population and that the obtained result can be biased towards the sample.

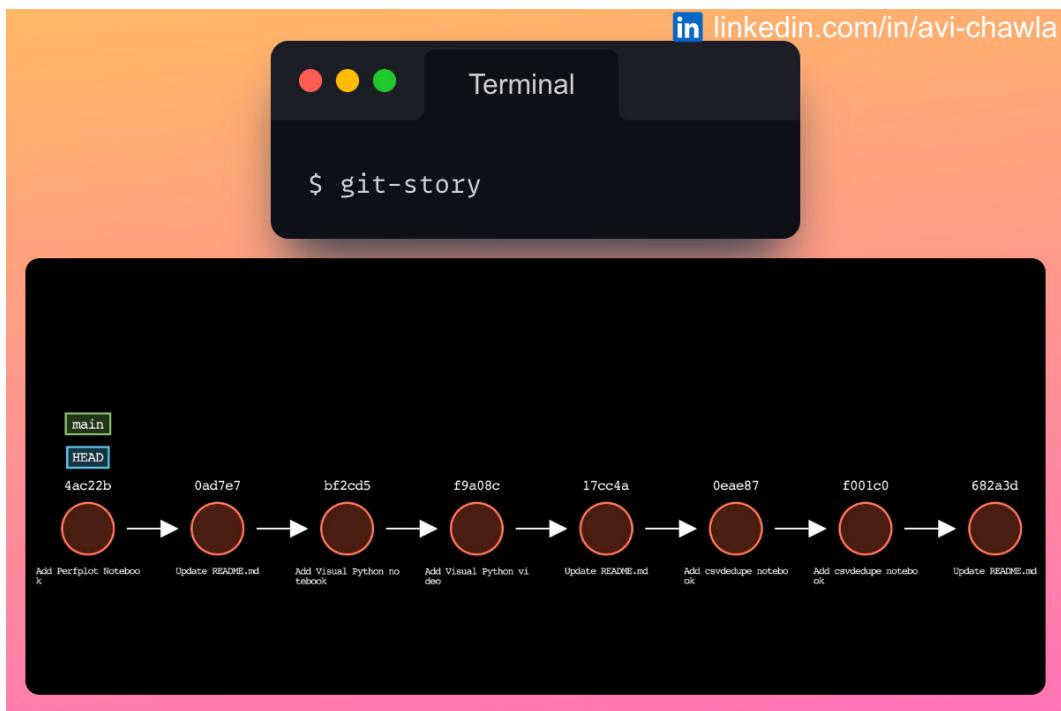
Thus, to generate an unbiased estimate, it uses  $(n-1)$  as the dividing factor instead of  $n$ . In statistics, this is also known as Bessel's correction.

NumPy, however, does not make any such correction.

Find more info here: [Bessel's correction](#).



# Visualize Commit History of Git Repo With Beautiful Animations



As the size of your project grows, it can get difficult to comprehend the Git tree.

Git-story is a command line tool to create elegant animations for your git repository.

It generates a video that depicts the commits, branches, merges, HEAD commit, and many more. Find more info in the comments.

Please watch a video version of this post here: [Video](#).

Read more: [Git-story](#).



# Perfplot: Measure, Visualize and Compare Run-time With Ease



Here's an elegant way to measure the run-time of various Python functions.

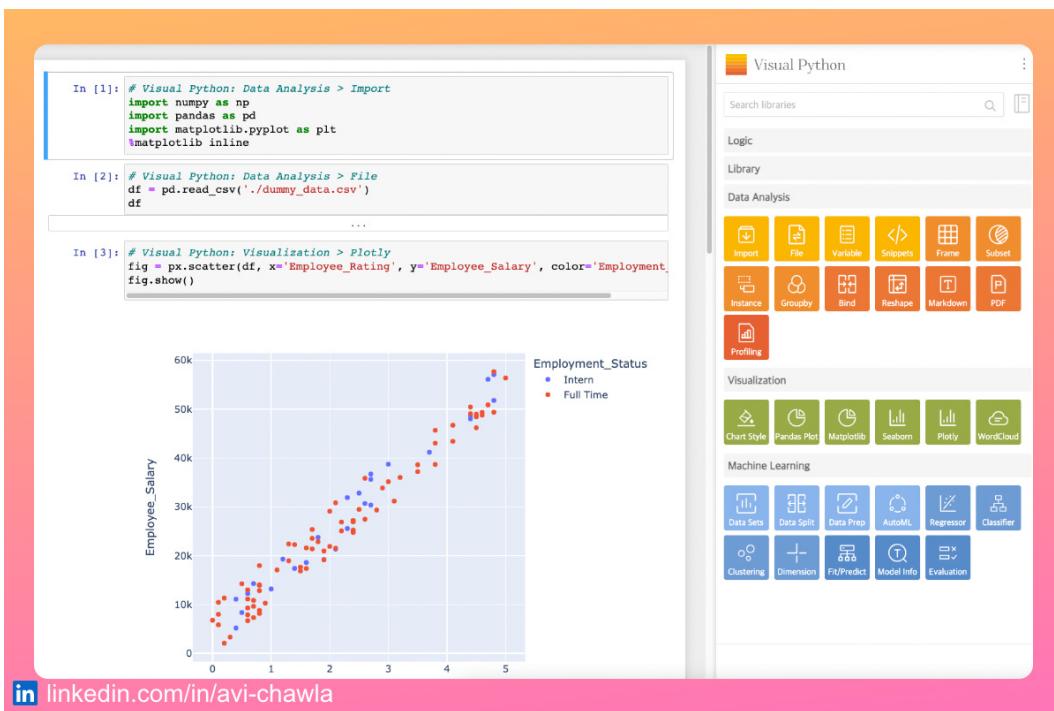
Perfplot is a tool designed for quick run-time comparisons of many functions/algorithms.

It extends Python's `timeit` package and allows you to quickly visualize the run-time in a clear and informative way.

Find more info: [Perfplot](#).



# This GUI Tool Can Possibly Save You Hours Of Manual Work



Please watch a video version of this post for better understanding: [Link](#).

This is indeed one of the coolest and most useful Jupyter notebook-based data science tools.

Visual Python is a GUI-based python code generator. Using this, you can easily eliminate writing code for many repetitive tasks. This includes importing libraries, I/O, Pandas operations, plotting, etc.

Moreover, with the click of a couple of buttons, you can import the code for many ML-based utilities. This covers sklearn models, evaluation metrics, data splitting functions, and many more.

Read more: [Visual Python](#).



# How Would You Identify Fuzzy Duplicates In A Data With Million Records?

	First_Name	Last_Name	Address	Phone
0	Daniel	Lopez	719 Greene St. East Rhonda	9371184929
1	Daniel	Nan	719 Green Street East Rhoda	93711-84929
2	Alan	Martin	982 Carol Harbors Apart.	7481919235
3	Alan Martin	Nan	982 Carol Aparments	748-191-9235
4	Philip	Owens	2578 Banks Ford	869-6922x9581
5	Shannon	White	USCGC Molina	(150)082-7982
6	Julia	Anderson	09162 Mason Mnts.	698-1590x3236
7	Juliya	Anderrson	9162 Mason Street Mountain	69815903236

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Data with fuzzy duplicates

Command Line

```
$ csvdedupe input.csv \
--field_names First_Name Last_Name Address Phone \
--output_file output.csv
```

Marked Duplicates



Cluster ID	First_Name	Last_Name	Address	Phone
0	Daniel	Lopez	719 Greene St. East Rhonda	9371184929
0	Daniel	nan	719 Green Street East Rhoda	93711-84929
1	Alan	Martin	982 Carol Harbors Apart.	7481919235
1	Alan Martin	nan	982 Carol Aparments	748-191-9235
2	Philip	Owens	2578 Banks Ford	869-6922x9581
3	Shannon	White	USCGC Molina	(150)082-7982
4	Julia	Anderson	09162 Mason Mnts.	698-1590x3236
4	Juliya	Anderrson	9162 Mason Street Mountain	69815903236

Imagine you have over a million records with fuzzy duplicates. How would you identify potential duplicates?

The naive approach of comparing every pair of records is infeasible in such cases. That's over  $10^{12}$  comparisons ( $n^2$ ). Assuming a speed of 10,000 comparisons per second, it will take roughly 3 years to complete.

The csvdedupe tool (linked in comments) solves this by cleverly reducing the comparisons. For instance, comparing the name "Daniel" to "Philip" or "Shannon" to "Julia" makes no sense. They are guaranteed to be distinct records.



Thus, it groups the data into smaller buckets based on rules. One rule could be to group all records with the same first three letters in the name.

This way, it drastically reduces the number of comparisons with great accuracy.

Read more: [csvdedupe](#).



# Stop Previewing Raw DataFrames. Instead, Use DataTables.

The screenshot shows a Jupyter Notebook cell with the following code:

```
In [1]: import pandas as pd  
from jupyter_datatables import init_datatables_mode  
  
In [2]: init_datatables_mode()  
  
In [3]: pd.read_csv("employee_dataset.csv")
```

Below the code, there is a search bar with the text "Search: andrade". The DataFrame preview includes columns: Name, Company\_Name, Employee\_Job\_Title, Employee\_City, and Employee\_Country. Each column has a yellow "OBJECT" button above it. Below the columns, there are five horizontal bar charts representing the distribution of data for each column. The preview shows 10 entries from a total of 69, with a sample size of 1,000 out of 1,000. A navigation bar at the bottom shows pages 1 through 7.

Out[3]: Sample size: 1,000 out of 1,000

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

After loading any dataframe in Jupyter, we preview it. But it hardly tells anything about the data.

One has to dig deeper by analyzing it, which involves simple yet repetitive code.

Instead, use [Jupyter-DataTables](#).

It supercharges the default preview of a DataFrame with many common operations. This includes sorting, filtering, exporting, plotting column distribution, printing data types, and pagination.

Please view a video version here for better understanding: [Post Link](#).



# 🚀 A Single Line That Will Make Your Python Code Faster

```
def func_without_numba():
    result = []
    for a in range(10000):
        for b in range(10000):
            if (a+b)%11 == 0:
                result.append((a,b))

func_without_numba()
# Run-time: 8.34 sec
```

~33x Faster

```
from numba import njit

@njit
def func_with_numba():
    # same code

func_with_numba()
# Run-time: 0.25 sec
```

[in](https://linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

If you are frustrated with Python's run-time, here's how a single line can make your code blazingly fast.

Numba is a just-in-time (JIT) compiler for Python. This means that it takes your existing python code and generates a fast machine code (at run-time).

Thus, post compilation, your code runs at native machine code speed. Numba works best on code that uses NumPy arrays and functions, and loops.

Get Started: [Numba Guide](#).



# Prettify Word Clouds In Python

```
from wordcloud import WordCloud

wc = WordCloud().generate(text)
```

```
from PIL import Image

mask = Image.open("pylogo.png")

wc = WordCloud(mask=mask,
               contour_width=4,
               ...)
wc.generate(text)
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

If you use word clouds often, here's a quick way to make them prettier.

In Python, you can easily alter the shape and color of a word cloud. By supplying a mask image, the resultant world cloud will take its shape and appear fancier.

Find more info here: [Notebook Link](#).



# How to Encode Categorical Features With Many Categories?

The screenshot shows a Jupyter Notebook cell with the following code:

```
import category_encoders as ce

enc = ce.BinaryEncoder(cols=['class'])

enc.fit_transform(data["class"])

    class_0  class_1  class_2
0        0        0        1
1        0        1        0
2        0        1        1
3        1        0        0
4        0        0        1
```

To the right of the code, there is a table labeled "data" with columns "gender" and "class". The data is as follows:

	gender	class
0	Male	A
1	Female	B
2	Male	C
3	Female	D
4	Female	A

A large curly brace on the right side of the matrix indicates that the three columns represent the binary encoding of the "class" column. The word "Binary" is written next to the brace.

We often encode categorical columns with one-hot encoding. But the feature matrix becomes sparse and unmanageable with many categories.

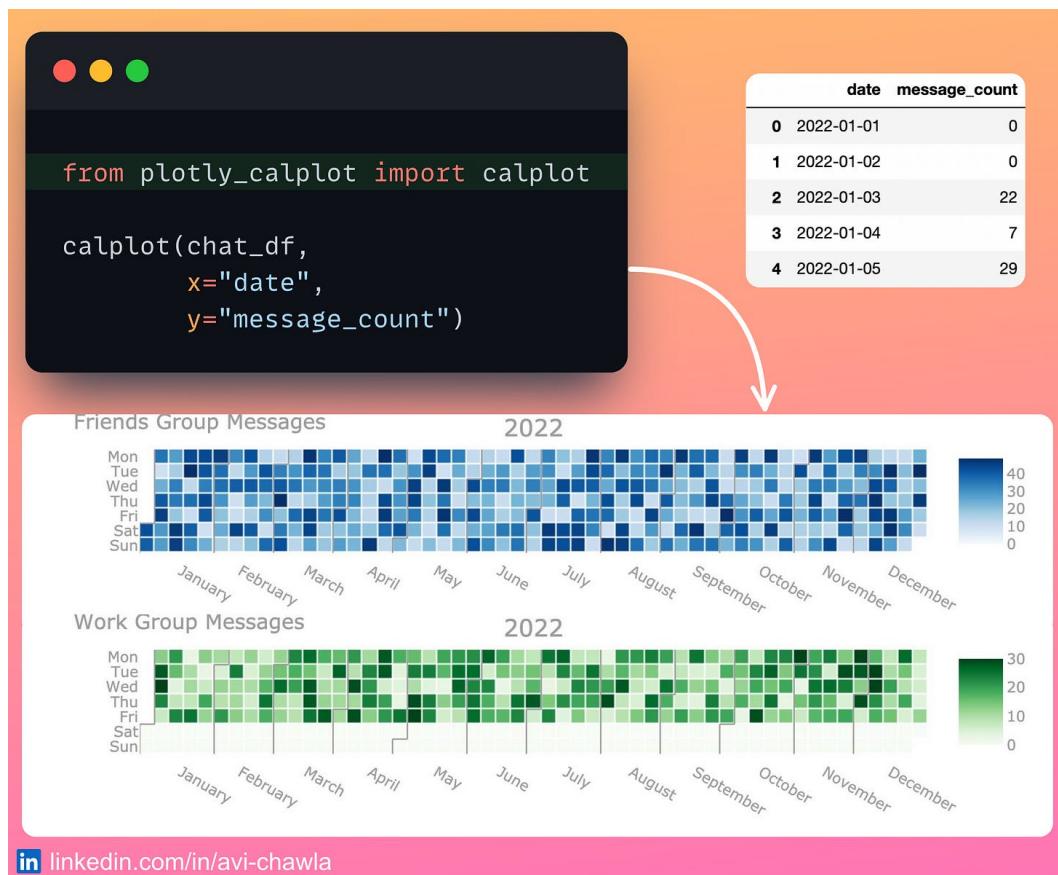
The category-encoders library provides a suite of encoders specifically for categorical variables. This makes it effortless to experiment with various encoding techniques.

For instance, I used its binary encoder above to represent a categorical column in binary format.

Read more: [Documentation](#).



# Calendar Map As A Richer Alternative to Line Plot



Ever seen one of those calendar heat maps? Here's how you can create one in two lines of Python code.

A calendar map offers an elegant way to visualize daily data. At times, they are better at depicting weekly/monthly seasonality in data instead of line plots. For instance, imagine creating a line plot for "Work Group Messages" above.

To create one, you can use "plotly\_calplot". Its input should be a DataFrame. A row represents the value corresponding to a date.

Read more: [Plotly Calplot](#).



# 10 Automated EDA Tools That Will Save You Hours Of (Tedious) Work

# 10 Automated EDA

# Tools That Will

# Save You Hours

# Of (Tedious) Work

Most steps in a data analysis task stay the same across projects. Yet, manually digging into the data is tedious and time-consuming, which inhibits productivity.

Here are 10 EDA tools that automate these repetitive steps and profile your data in seconds.

Please find this full document in my LinkedIn post: [Post Link](#).



# Why KMeans May Not Be The Apt Clustering Algorithm Always

The slide is divided into two main sections: 'Incorrect clusters' (top) and 'Correct clusters' (bottom).  
The 'Incorrect clusters' section shows a scatter plot of data points forming two crescent-shaped clusters. A green line represents the cluster boundaries predicted by KMeans, which fails to correctly separate the two classes. The 'Prediction' legend indicates blue dots for 'Class A' and red dots for 'Class B'.  
The 'Correct clusters' section shows the same data points, but the DBSCAN algorithm has correctly identified two distinct clusters, represented by blue and red dots.  
On the left, there are two code snippets:  
Top: `from sklearn import cluster  
cluster.KMeans(2).fit(X)`  
Bottom: `from sklearn import cluster  
cluster.DBSCAN().fit(X)`  
A LinkedIn link is also present at the bottom left: [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla).

KMeans is a popular clustering algorithm. Yet, its limitations make it inapplicable in many cases.

For instance, KMeans clusters the points purely based on locality from centroids. Thus, it can create wrong clusters when data points have arbitrary shapes.

Among the many possible alternatives is DBSCAN, which is a density-based clustering algorithm. Thus, it can identify clusters of arbitrary shape and size.

This makes it robust to data with non-spherical clusters and varying densities. Find more info in the comments.

Find more here: [Sklearn Guide](#).



# Converting Python To LaTeX Has Possibly Never Been So Simple

```
import latexify
import math

@latexify.function      Add decorator
def roots(a, b, c):
    return (-b + math.sqrt(b**2 - 4*a*c)) / (2*a)
```

roots

$$\text{roots}(a, b, c) = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

```
@latexify.function
def fib(n):
    if n<2:
        return 1
    else:
        return fib(n-1) + fib(n-2)
```

fib

$$\text{fib}(n) = \begin{cases} 1, & \text{if } n < 2 \\ \text{fib}(n - 1) + \text{fib}(n - 2), & \text{otherwise} \end{cases}$$

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

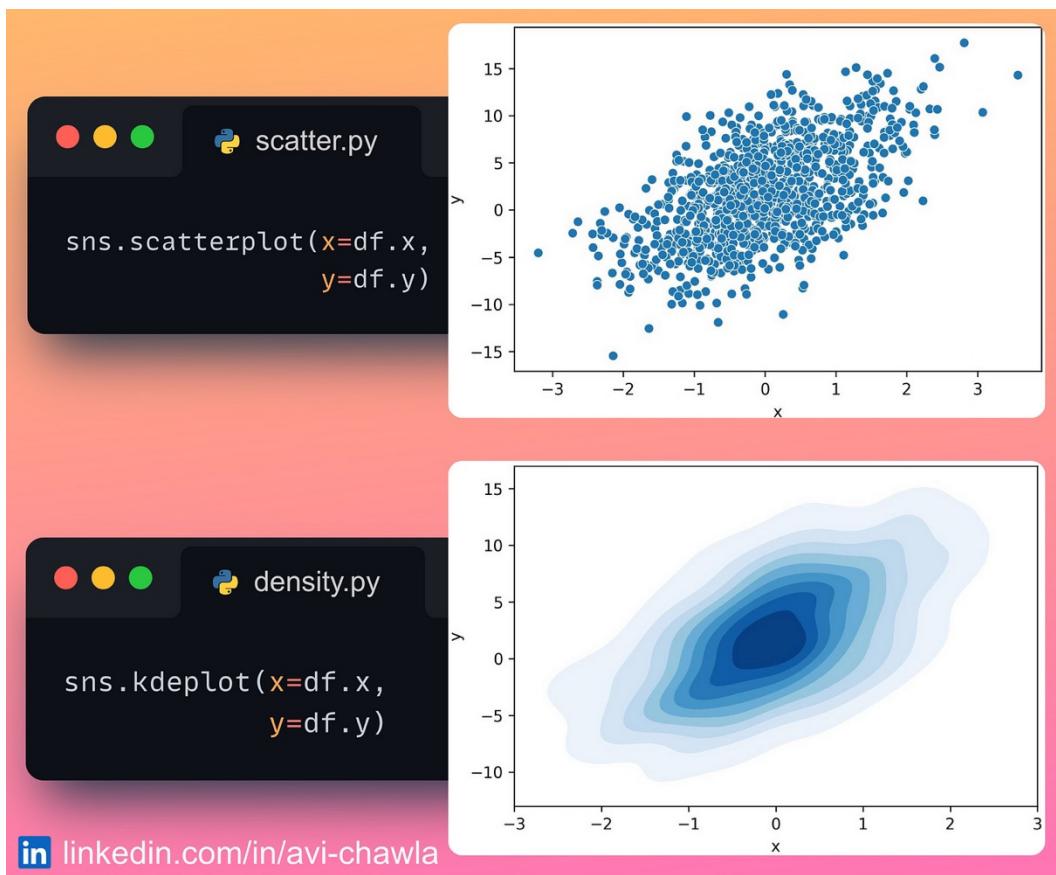
If you want to display python code and its output as LaTeX, try `latexify_py`. With this, you can print python code as a LaTeX expression and make your code more interpretable.

What's more, it can also generate LaTeX code for python code. This saves plenty of time and effort of manually writing the expressions in LaTeX.

Find more info here: [Repository](#).



# Density Plot As A Richer Alternative to Scatter Plot



Scatter plots are extremely useful for visualizing two sets of numerical variables. But when you have, say, thousands of data points, scatter plots can get too dense to interpret.

A density plot can be a good choice in such cases. It depicts the distribution of points using colors (or contours). This makes it easy to identify regions of high and low density.

Moreover, it can easily reveal clusters of data points that might not be obvious in a scatter plot.

Read more: [Docs](#).



# 30 Python Libraries to (Hugely) Boost Your Data Science Productivity

# 30 Python Libraries to (Hugely) Boost Your Data Science Productivity

Here's a collection of 30 essential open-source data science libraries. Each has its own use case and enormous potential to skyrocket your data science skills.

I would love to know the ones you use.

Please find this full document in my LinkedIn post: [Post Link](#).



# Sklearn One-liner to Generate Synthetic Data

```
from sklearn.datasets import make_classification

## create data
X, y = make_classification(n_samples=50,
                           n_features=4,
                           n_classes=2)

>>> print(X)
array([[-0.36,  1.01,  0.19, -1.18],
       [-0.29,  1.21,  0.22, -1.92],
       ...
       [-2.12,  1.82,  0.59,  3.18]])

>>> print(y)
array([0, 1, ..., 0, 1])
```

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Often for testing/building a data pipeline, we may need some dummy data.

With Sklearn, you can easily create a dummy dataset for regression, classification, and clustering tasks.

More info here: [Sklearn Docs](#).



# Label Your Data With The Click Of A Button

```
In [3]: from ipyannotate import annotate
from ipyannotate.buttons import ValueButton as Button

In [5]: annotation = annotate(images_data,
                             buttons=[Button('Dog'), # label buttons list
                                       Button('Cat')])
annotation
```

Dog Cat

```
In [6]: labels = [task.value for task in annotation.tasks] # get labels
labels
out[6]: ['Cat', 'Dog', 'Dog', 'Cat']
```

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Often with unlabeled data, one may have to spend some time annotating/labeling it.

To do this quickly in a jupyter notebook, use **ipyannotate**. With this, you can annotate your data by simply clicking the corresponding button.

Read more: [ipyannotate](#).

Watch a video version of this post on LinkedIn: [Post Link](#).



# Analyze A Pandas DataFrame Without Code

The screenshot shows a Pandas GUI application window. At the top, there are three code cells:

```
In [1]: import pandas as pd  
from pandasgui import show  
  
In [2]: df = pd.read_csv("Dummy_Dataset.csv")  
  
In [4]: show(df)
```

Below the code cells is a DataFrame viewer with the following columns:

Index	Name	Company_Name	Employee_Job_Title	Employee_City	Employee_Country
0	Michael Clark	James and Sons	Regulatory affairs officer	Ricardomouth	Western Sahara
1	Edwin Smith	Baker, Allen and Edwards	Trading standards officer	Whitakerbury	Singapore
2	Leslie Donovan	Nelson-Li	Naval architect	New Russellton	Niue
3	Phyllis King	Taylor-Ramos	Make	Ricardomouth	Tokelau
4	Joshua Patterson	Thomas-Spencer	Retail merchandiser	Wandfort	Croatia
5	Cheyenne Torres	Nelson-Li	Actuary	Kristaburgh	Thailand
6	Jonathan Chen	Wallace, Smith and Shepard	Actuary	New Cindychester	Equatorial Guinea
7	Scott Powell	Scott Inc	Administrator	Ricardomouth	Palau
8	Theresa Doyle	Bullock-Carrillo	Production engineer	Ricardomouth	Cape Verde
9	Kristen Harrington	James and Sons	Sales promotion account executive	Kristaburgh	United States of America
10	Joseph Hunt	Wallace, Smith and Shepard	Armed forces logistics/support/administrative officer	Aliciafort	Equatorial Guinea
11	Tracy King	Bullock-Carrillo	Garnett/Textile technologist	New Cindychester	Nepal
12	Julie Moses	Bullock-Carrillo	Energy manager	North Melissafurt	Ghana
13	Tanya Cross	James and Sons	Actuary	Ricardomouth	Botswana
14	Christopher Berry	Marshall-Holloway	Actuary	Ricardomouth	Norway
15	Rebecca Jimenez	White, McClain and Cobb	Diplomatic Services operational officer	West Jamesview	Bangladesh
16	Adam Sampson	James and Sons	Investment banker/corporate	West Jamesview	Honduras
17	Austin Smith	Marshall-Holloway	Administrator	New Russellton	Thailand
18	John Edwards	Taylor-Ramos	Actuary	New Russellton	Gibraltar
19	Theresa Espinoza	James and Sons	Energy manager	New Cindychester	Guyana
20	Lisa McCall	Bullock-Carrillo	Ergonomist	West Jamesview	Saint Pierre and Miquelon
21	Scott Escobar	Nichols-James	Trading standards officer	Kristaburgh	Kyrgyz Republic
22	Christopher Callahan	Thomas-Spencer	Trading standards officer	North Melissafurt	Mali
23	Tara Shepard	Bullock-Carrillo	Naval architect	Ricardomouth	Jamaica

On the left side of the window, there are two panels: "Filters" and "Columns". The "Filters" panel has a search bar and a message "No filters defined". The "Columns" panel lists all the columns of the DataFrame.

At the bottom left, there is a LinkedIn icon followed by the URL [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla).

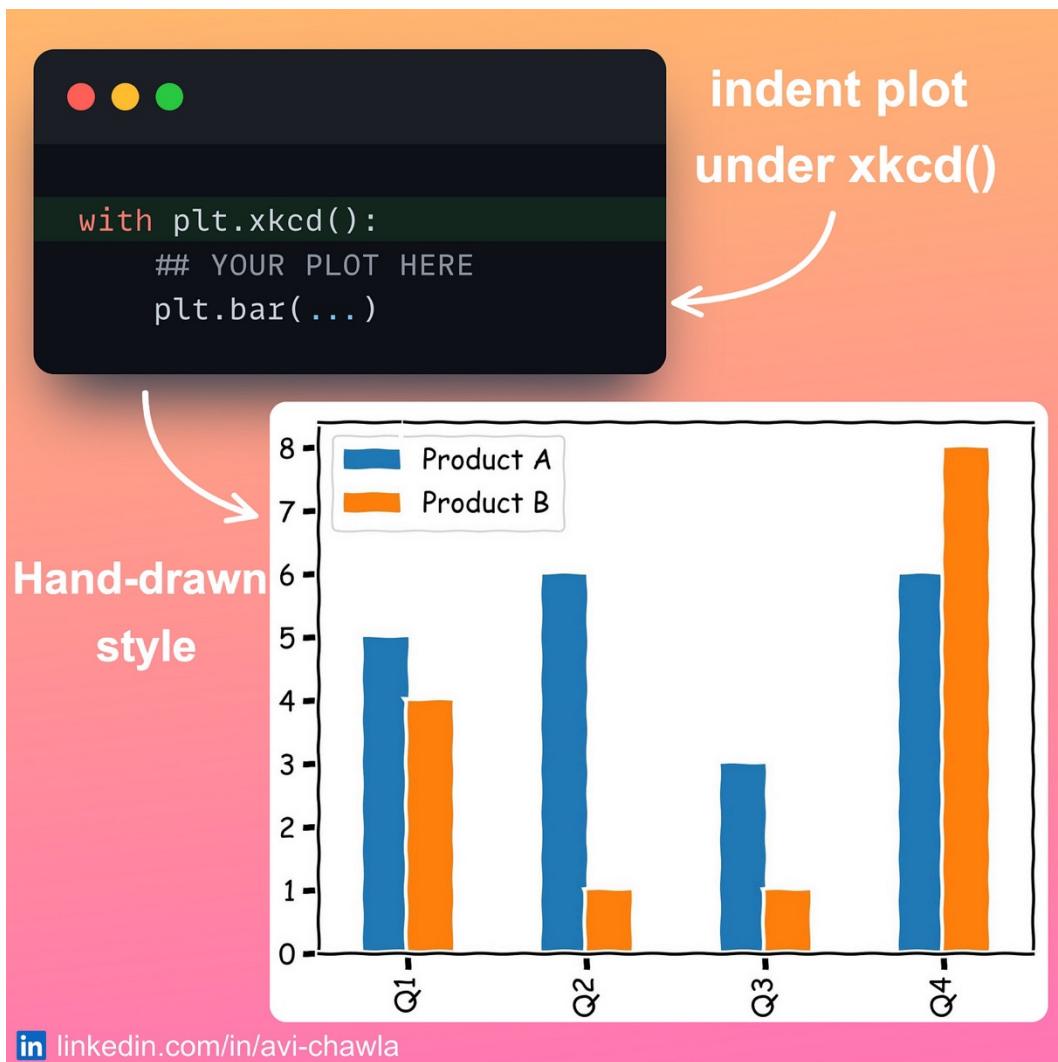
If you want to analyze your dataframe in a GUI-based application, try Pandas GUI. It provides an elegant GUI for viewing, filtering, sorting, describing tabular datasets, etc.

What's more, using its intuitive drag-and-drop functionality, you can easily create a variety of plots and export them as code.

Watch a video version of this post on LinkedIn: [Post Link](#).



# Python One-Liner To Create Sketchy Hand-drawn Plots



[xkcd](#) comic is known for its informal and humorous style, as well as its stick figures and simple drawings.

Creating such visually appealing hand-drawn plots is pretty simple using matplotlib. Just indent the code in a `plt.xkcd()` context to display them in comic style.

Do note that this style is just used to improve the aesthetics of a plot through hand-drawn effects. However, it is not recommended for formal presentations, publications, etc.

Read more: [Docs](#).



# 70x Faster Pandas By Changing Just One Line of Code

The image shows a Mac desktop with two terminal windows side-by-side. The background is a gradient from orange to red.

**Pandas.py Terminal:**

```
import pandas as pd

data = "file.csv" ## 2M Rows

df = pd.read_csv(data)
## 3.6 sec

pd.concat([df for _ in range(20)])
## 7.1 sec
```

**Modin.py Terminal:**

```
import modin.pandas as pd

data = "file.csv" ## 2M Rows

df = pd.read_csv(data)
## 1.3 sec (2.75x Faster)

pd.concat([df for _ in range(20)])
## 0.1 sec (70x Faster)
```

A white arrow points from the text "Modify Import Statement" on the left towards the "Modin.py" terminal window. Another white arrow points from the text "7 GB Dataset" on the right towards the "Pandas.py" terminal window.

**LinkedIn:** [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

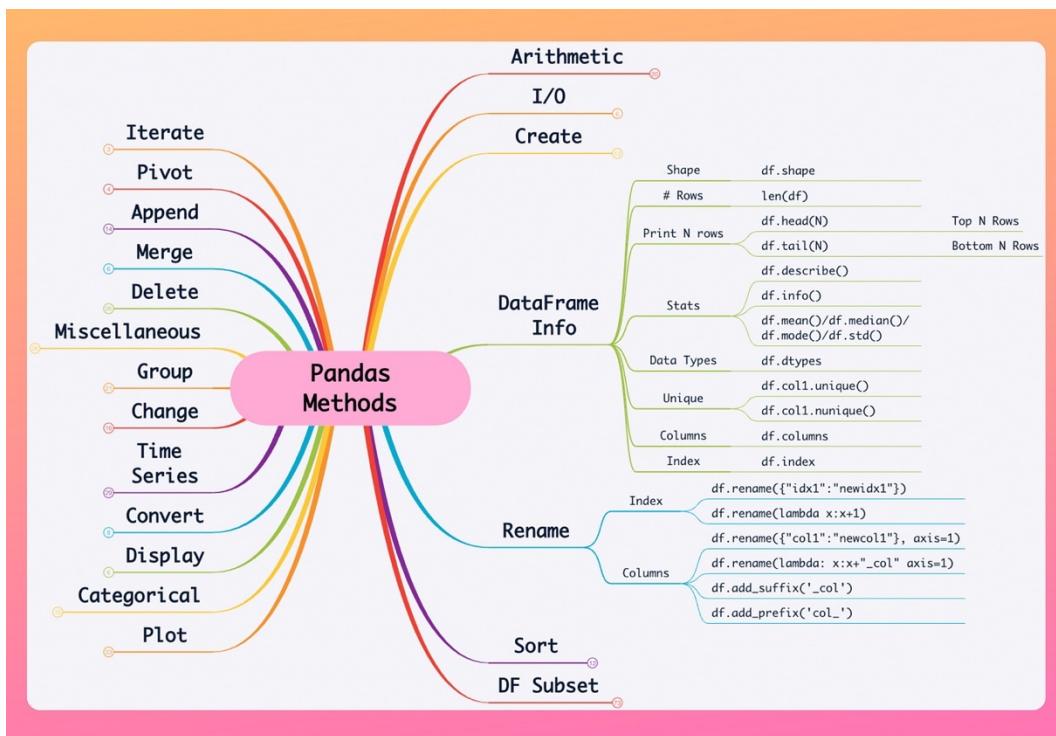
It is challenging to work on large datasets in Pandas. This, at times, requires plenty of optimization and can get tedious as the dataset grows further.

Instead, try Modin. It delivers instant improvements with no extra effort. Change the import statement and use it like the Pandas API, with significant speedups. Find more info in the comments.

Read more: [Modin Guide](#).



# An Interactive Guide To Master Pandas In One Go



Here's a mind map illustrating Pandas Methods on one page. How many do you know :)

- ◆ Load/Save
- ◆ DataFrame info
- ◆ Filter
- ◆ Merge
- ◆ Time-series
- ◆ Plot
- ◆ and many more, in a single map.

Find the full diagram here: [Pandas Mind Map.](#)



# Make Dot Notation More Powerful in Python

```
class Square:
    def __init__(self, length):
        self._side = length

    @property → Getter
    def side(self):
        return self._side

    @side.setter → Setter
    def side(self, length):
        if length<0:
            raise ValueError("Side cannot be negative")
        else:
            self._side = length
```

Raises errors during assignment

```
>>> s = Square(10)
>>> s.side # Getter
10
>>> s.side = -2 # Setter (with dot)
ValueError: Side cannot be negative
```

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Dot notation offers a simple and elegant way to access and modify the attributes of an instance.

Yet, it is a good programming practice to use the getter and setter method for such purposes. This is because it offers more control over how attributes are accessed/changed.

To leverage both in Python, use the **@property** decorator. As a result, you can use the dot notation and still have explicit control over how attributes are accessed/set.



# The Coolest Jupyter Notebook Hack

The diagram illustrates three methods to access the output of a previously run Jupyter cell. It features a vertical gradient background transitioning from orange at the top to pink at the bottom. Three numbered arrows point from the text labels to their corresponding code examples:

- 1**: Points to In [3]: `_2` and Out[3]: `array([1, 2, 3])`.
- 2**: Points to In [4]: `Out[2]` and Out[4]: `array([1, 2, 3])`.
- 3**: Points to In [5]: `_oh[2]` and Out[5]: `array([1, 2, 3])`.

**In [1]:** `import numpy as np`

**In [2]:** `np.array([1,2,3])`

**Out[2]:** `array([1, 2, 3])`

**In [3]:** `_2`

**Out[3]:** `array([1, 2, 3])`

**In [4]:** `Out[2]`

**Out[4]:** `array([1, 2, 3])`

**In [5]:** `_oh[2]`

**Out[5]:** `array([1, 2, 3])`

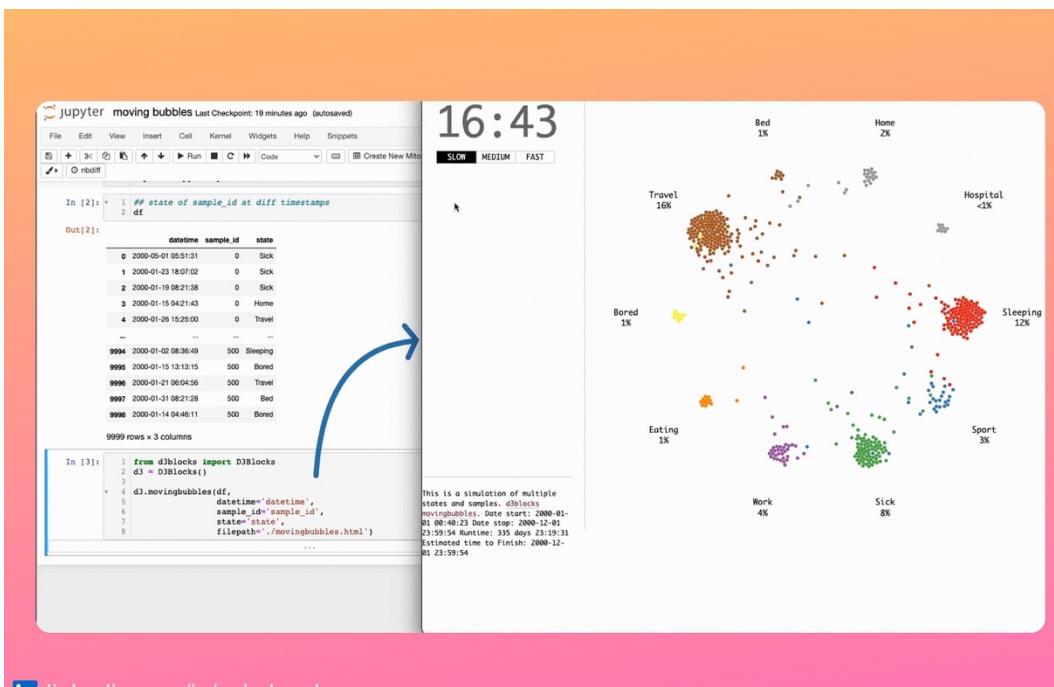
[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Have you ever forgotten to assign the results to a variable in Jupyter? Rather than recomputing the result by rerunning the cell, here are three ways to retrieve the output.

- 1) Use the underscore followed by the output-cell-index.
- 2/3) Use the **Out** or **\_oh** dict and specify the output-cell-index as the key.



# Create a Moving Bubbles Chart in Python



[in](https://linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Ever seen one of those moving points charts? Here's how you can create one in Python in just three lines of code.

A Moving Bubbles chart is an elegant way to depict the movements of entities across time. Using this, we can easily determine when clusters appear in our data and at what state(s).

To create one, you can use "[d3blocks](#)". Its input should be a DataFrame. A row represents the state of a sample at a particular timestamp.



# Skorch: Use Scikit-learn API on PyTorch Models

**Define Pytorch model**

```
class MyModel(nn.Module):
    def __init__(self):
        ## Define Network

    def forward(self, x):
        ## Forward Pass
```

**Use Scikit-learn API on model**

```
from skorch import NeuralNetClassifier

model = NeuralNetClassifier(
    MyModel,
    lr=0.1,
    criterion=nn.MSELoss
)

model.fit(X, y)
preds = model.predict(X)
```

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

skorch is a high-level library for PyTorch that provides full Scikit-learn compatibility. In other words, it combines the power of PyTorch with the elegance of sklearn.

Thus, you can train PyTorch models in a way similar to Scikit-learn, using functions such as fit, predict, score, etc.

Using skorch, you can also put a PyTorch model in the sklearn pipeline, and many more.

Overall, it aims at being as flexible as PyTorch while having a clean interface as sklearn.

Read more: [Documentation](#).



# Reduce Memory Usage Of A Pandas DataFrame By 90%

The screenshot shows a Jupyter Notebook interface with two code cells and a data preview.

**Code Cell 1:**

```
## df.shape: (10^7, 2)

>>> df.A.dtype
dtype('int64')

## Range: [-2^63, 2^63-1]

>>> df.A.min(), df.A.max()
(1, 100)

>>> df.A.memory_usage()
76.3 MB
```

**Code Cell 2:**

```
df["A"] = df.A.astype(np.int8)
## Range: [-128, 127]

>>> df.A.memory_usage()
9.5 MB # (~90% Lower)
```

**Data Preview:**

	A	B
0	38	46
1	28	58
2	47	82
3	88	87
4	13	78

**Annotations:**

- Two arrows point from the text "Supported range larger than required" to the line `## Range: [-2^63, 2^63-1]`.
- An arrow points from the text "Convert to smaller datatype" to the line `df["A"] = df.A.astype(np.int8)`.

**LinkedIn Link:** [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

By default, Pandas always assigns the highest memory datatype to its columns. For instance, an integer-valued column always gets the int64 datatype, irrespective of its range.

To reduce memory usage, represent it using an optimized datatype, which is enough to span the range of values in your columns.

Read [this blog](#) for more info. It details many techniques to optimize the memory usage of a Pandas DataFrame.



# An Elegant Way To Perform Shutdown Tasks in Python

```
my_file.py
```

```
import atexit
@atexit.register
def final_function():
    print("COMPLETED EXECUTION!")

for i in range(5):
    print(f"num = {i}")
```

Add decorator to method

The decorator invokes the function at the end

```
Terminal
```

```
$ python my_file.py
```

```
num = 0
num = 1
num = 2
num = 3
num = 4
```

```
COMPLETED EXECUTION!
```

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Often towards the end of a program's execution, we run a few basic tasks such as saving objects, printing logs, etc.

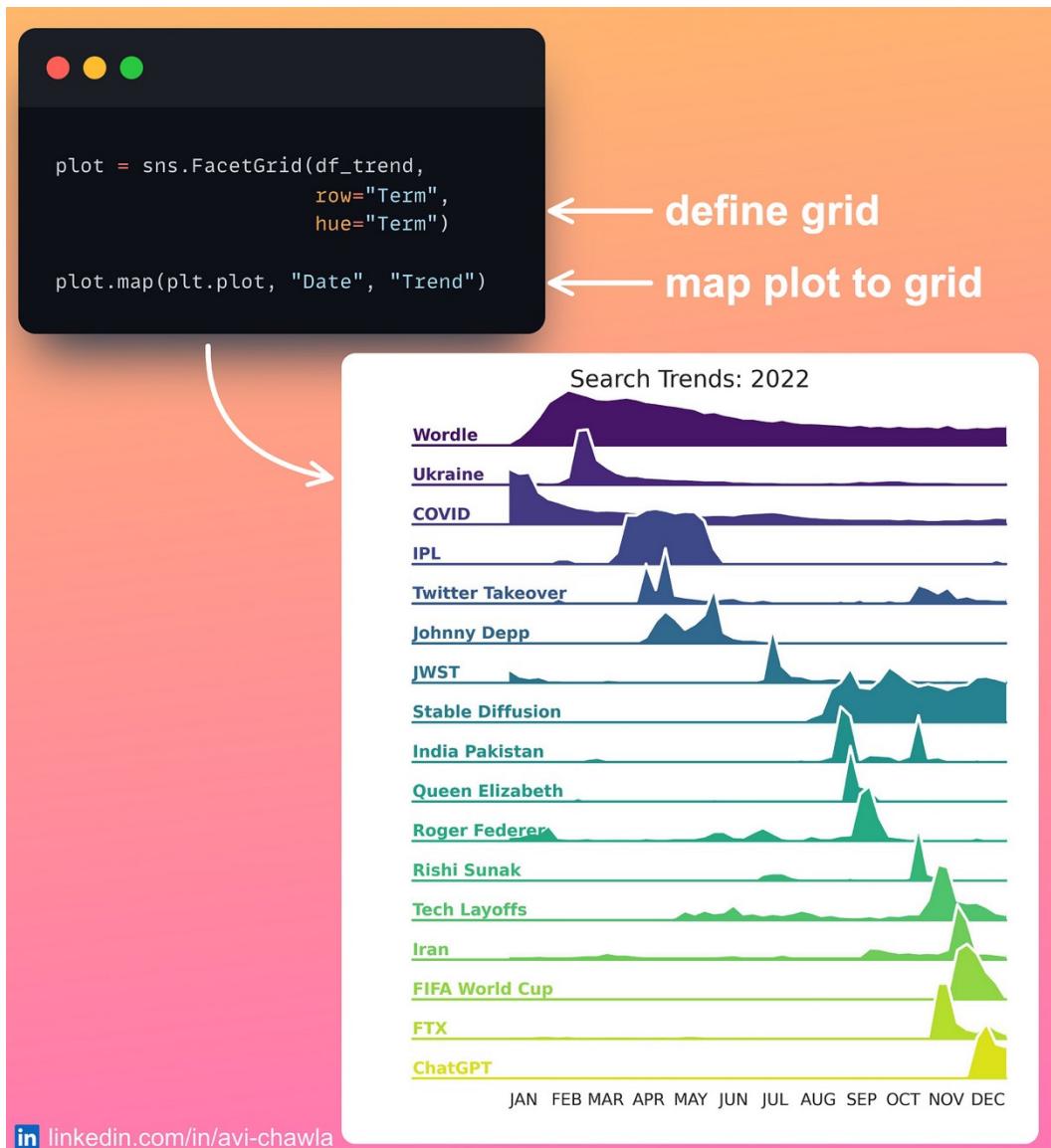
To invoke a method right before the interpreter is shutting down, decorate it with the `@atexit.register` decorator.

The good thing is that it works even if the program gets terminated unexpectedly. Thus, you can use this method to save the state of the program or print any necessary details before it stops.

Read more: [Documentation](#).



# Visualizing Google Search Trends of 2022 using Python



If your data has many groups, visualizing their distribution together can create cluttered plots. This makes it difficult to visualize the underlying patterns.

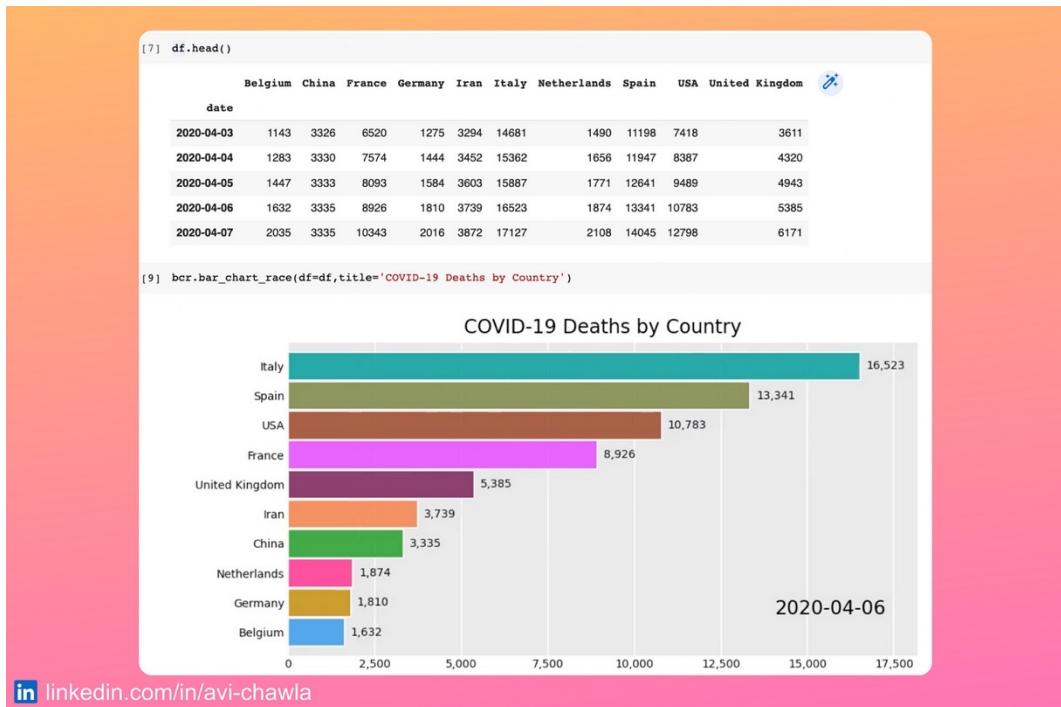
Instead, consider plotting the distribution across individual groups using FacetGrid. This allows you to compare the distributions of multiple groups side by side and see how they vary.

As shown above, a FacetGrid allows us to clearly see how different search terms trended across 2022.

P.S. I used the [year-in-search-trends](#) repository to fetch the trend data.



# Create A Racing Bar Chart In Python



Ever seen one of those racing bar charts? Here's how you can create one in Python in just two lines of code.

A racing bar chart is typically used to depict the progress of multiple values over time.

To create one, you can use the "**bar-chart-race**" library.

Its input should be a Pandas DataFrame where every row represents a single timestamp. The column holds the corresponding values for a particular category.

Read more: [Documentation](#).



# Speed-up Pandas Apply 5x with NumPy

Pandas Apply

```
def assign_class(num):
    if num<10:
        return "Class A"
    if num<50:
        return "Class B"
    return "Class C"

df.A.apply(assign_class)
## 1.02 s ± 20.5 ms per loop
```

A B C D

	A	B	C	D
0	19	80	39	36
1	20	97	47	9
2	3	63	16	69
3	68	20	58	37
4	63	71	51	32

10<sup>7</sup> rows

NumPy Select

```
condlist = [ df["A"]<10 , df["A"]<50 ]
resultlist = [ "Class A" , "Class B" ]

np.select(condlist, resultlist, "Class C")
## 0.20 s ± 7.14 ms per loop
```

~5x Faster

Default

linkedin.com/in/avi-chawla

While creating conditional columns in Pandas, we tend to use the **apply()** method almost all the time.

However, **apply()** in Pandas is nothing but a glorified for-loop. As a result, it misses the whole point of vectorization.

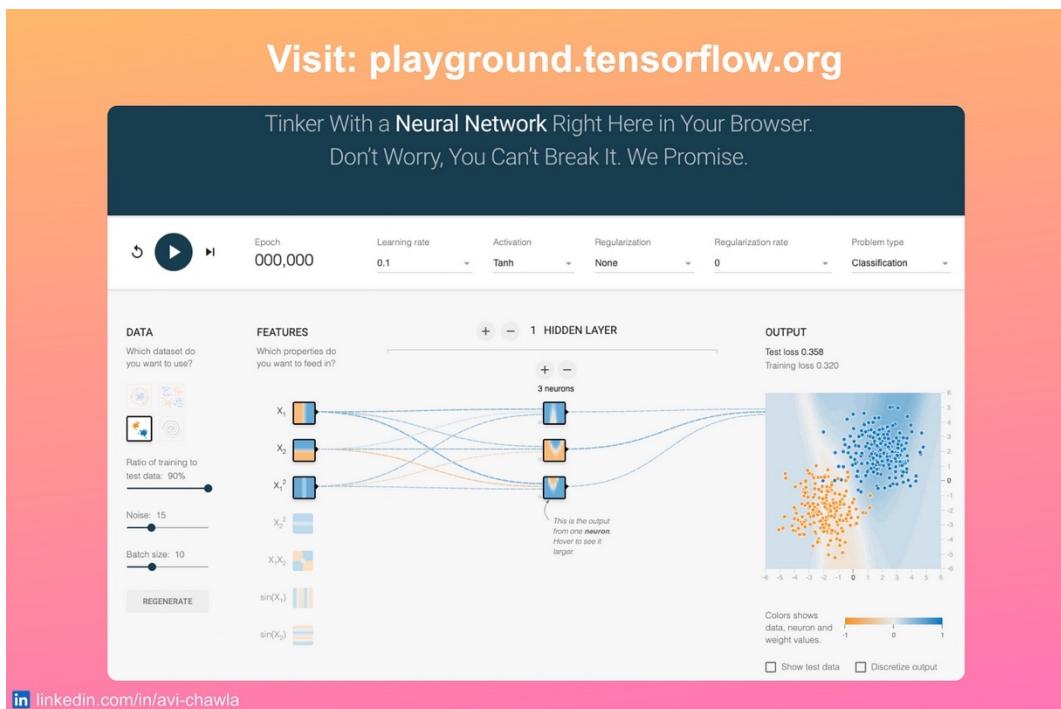
Instead, you should use the **np.select()** method to create conditional columns. It does the same job but is extremely fast.

The conditions and the corresponding results are passed as the first two arguments. The last argument is the default result.

Read more here: [NumPy docs](#).



# A No-Code Online Tool To Explore and Understand Neural Networks



Neural networks can be intimidating for beginners. Also, experimenting programmatically does not provide enough intuitive understanding about them.

Instead, try TensorFlow Playground. Its elegant UI allows you to build, train and visualize neural networks without any code.

With a few clicks, one can see how neural networks work and how different hyperparameters affect their performance. This makes it especially useful for beginners.

Try here: [Tensorflow Playground](https://playground.tensorflow.org).



# What Are Class Methods and When To Use Them?

```
class Rectangle:
    def __init__(self, width, height):
        self.width = width
        self.height = height

    @classmethod
    def from_square(cls, size):
        return Rectangle(size, size)
```

Define classmethod

create object using classmethod →

```
rect = Rectangle.from_square(5)

print(rect.width) # Output: 5
print(rect.height) # Output: 5
```

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Class methods, as the name suggests, are bound to the class and not the instances of a class. They are especially useful for providing an alternative interface for creating instances.

Moreover, they can be also used to define utility functions that are related to the class rather than its instances.

For instance, one can define a class method that returns a list of all instances of the class. Another use could be to calculate a class-level statistic based on the instances.

To define a class method in Python, use the **@classmethod** decorator. As a result, this method can be called directly using the name of the class.



# Make Sklearn KMeans 20x times faster

The image shows two terminal windows side-by-side. The top window is titled 'sklearn.py' and contains Python code for training a KMeans model with 8 clusters on a dataset 'x\_train' (shape: 500000, 1024). The bottom window is titled 'faiss.py' and contains Python code for training a KMeans model using the Faiss library with the same parameters. A large white arrow points from the 'faiss.py' window to the left, where the text '~20x Faster' is displayed in white against a pink background.

```
from sklearn.cluster import KMeans

kmeans = KMeans(8).fit(x_train)
# Training Time: 162s
```

x\_train shape: (500000, 1024)

```
import faiss

kmeans = faiss.Kmeans(d=1024, k=8)
kmeans.train(x_train)
# Training Time: 7.8s
```

~20x Faster

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

The KMeans algorithm is commonly used to cluster unlabeled data. But with large datasets, scikit-learn takes plenty of time to train and predict.

To speed-up KMeans, use Faiss by Facebook AI Research. It provides faster nearest-neighbor search and clustering.

Faiss uses "Inverted Index", an optimized data structure to store and index the data points. This makes performing clustering extremely efficient.

Additionally, Faiss provides parallelization and GPU support, which further improves the performance of its clustering algorithms.

Read more: [GitHub](#).



# Speed-up NumPy 20x with Numexpr

```
import numpy as np
import numexpr as ne
```

```
a = np.random.random(10**7)
b = np.random.random(10**7)
```

```
%timeit np.cos(a) + np.sin(b)
```

142 ms ± 257 µs per loop

```
%timeit ne.evaluate("cos(a) + sin(b)")
```

32.5 ms ± 229 µs per loop **~5x Faster**

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Numpy already offers fast and optimized vectorized operations. Yet, it does not support parallelism. This provides further scope for improving the run-time of NumPy.

To do so, use Numexpr. It allows you to speed up numerical computations with multi-threading and just-in-time compilation.

Depending upon the complexity of the expression, the speed-ups can range from 0.95x and 20x. Typically, it is expected to be 2x-5x.

Read more: [Documentation](#).



# A Lesser-Known Feature of Apply Method In Pandas

The screenshot shows a Jupyter Notebook interface with three code cells and a data preview.

- Code Cell 1:** A function definition:

```
def min_max(row):
    return max(row), min(row)
```
- Code Cell 2:** An application of the function to a DataFrame:

```
>>> df.apply(min_max, axis = 1)
0    (3, 1)
1    (6, 3)
```
- Code Cell 3:** An application of the function with a specific result type:

```
>>> df.apply(min_max, axis = 1,
            result_type="expand")
      0   1
0   3   1
1   6   3
```

Annotations highlight the output of the second cell as a "Pandas Series of Tuple" and the third cell as a "Pandas DataFrame". A LinkedIn link is also present at the bottom.

Pandas Series of Tuple

Pandas DataFrame

linkedin.com/in/avi-chawla

After applying a method on a DataFrame, we often return multiple values as a tuple. This requires additional steps to project it back as separate columns.

Instead, with the `result_type` argument, you can control the shape and output type. As desired, the output can be either a DataFrame or a Series.



# An Elegant Way To Perform Matrix Multiplication

```
import numpy as np

x = np.matmul(a, np.matmul(b, c))
```

```
x = a @ b @ c
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Matrix multiplication is a common operation in machine learning. Yet, chaining repeated multiplications using **matmul** function makes the code cluttered and unreadable.

If you are using NumPy, you can instead use the **@** operator to do the same.



# Create Pandas DataFrame from Dataclass

The diagram illustrates the process of creating a Pandas DataFrame from a Dataclass. It consists of two code snippets presented in a dark-themed window, each with a title and a callout arrow.

**define dataclass**

```
from dataclasses import dataclass

@dataclass
class Point:
    x_loc:int
    y_loc:int
```

**list of dataclass objects**

```
points = [Point(5, 5),
          Point(1, 4),
          Point(2, 3)]

pd.DataFrame(points)
"""
   x_loc  y_loc
0      5      5
1      1      4
2      2      3
"""
```

**in** [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

A Pandas DataFrame is often created from a Python list, dictionary, by reading files, etc. However, did you know you can also create a DataFrame from a Dataclass?

The image demonstrates how you can create a DataFrame from a list of dataclass objects.



# Hide Attributes While Printing A Dataclass Object

The slide features two terminal windows side-by-side. The left terminal shows the default behavior where all attributes are printed, while the right terminal shows a modified class definition where one attribute is hidden from the print output.

**Terminal 1 (Left): Prints all attributes**

```
from dataclasses import dataclass

@dataclass
class Student:
    name:str
    key:str

Jane = Student("Jane", "27HD")

print(Jane)
Student(name='Jane', key='27HD')
```

**Terminal 2 (Right): Attribute hidden in print**

```
from dataclasses import dataclass, field

@dataclass
class Student:
    name:str
    key:str = field(repr = False)

Jane = Student("Jane", "27HD")

print(Jane)
Student(name='Jane')
```

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

By default, a dataclass prints all the attributes of an object declared during its initialization.

But if you want to hide some specific attributes, declare **repr=False** in its field, as shown above.



# List : Tuple :: Set : ?

A set cannot be added as a key

```
my_set = {1, 2, 3}

my_dict = {my_set: "A set"}
## TypeError: unhashable type: 'set'
```

Use  
frozenset

```
## frozenset
my_set = frozenset({1, 2, 3})

my_dict = {my_set: "A frozen set"}

my_dict[my_set]
"A frozen set"
```

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Dictionaries in Python require their keys to be immutable. As a result, a set cannot be used as keys as it is mutable.

Yet, if you want to use a set, consider declaring it as a frozenset.

It is an immutable set, meaning its elements cannot be changed after it is created. Therefore, they can be safely used as a dictionary's key.



# Difference Between Dot and Matmul in NumPy

The image shows two Jupyter notebook cells side-by-side. The left cell, titled 'dot.py', contains the following code:

```
>>> a:np.array # Shape: (a,b,c,d)
>>> b:np.array # Shape: (p,q,d,r)
```

Annotations for this cell:

- **a\*b\*c VECTORS of shape (d)** points to the first line of code.
- **p\*q\*r VECTORS of shape (d)** points to the second line of code.
- **dot product of a\*b\*c and p\*q\*r VECTORS** points to the result of the np.dot(a, b) call.

The right cell, titled 'matmul.py', contains the following code:

```
>>> a:np.array # Shape: (a,b,c,d)
>>> b:np.array # Shape: (a,b,d,e)
```

Annotations for this cell:

- **a\*b MATRICES of shape (c,d)** points to the first line of code.
- **a\*b MATRICES of shape (d,e)** points to the second line of code.
- **Matrix product of a\*b MATRICES** points to the result of the np.matmul(a, b) call.

At the bottom left of the image is a LinkedIn link: [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla).

The **np.matmul()** and **np.dot()** methods produce the same output for 2D (and 1D) arrays. This makes many believe that they are the same and can be used interchangeably, but that is not true.

The **np.dot()** method revolves around individual vectors (or 1D arrays). Thus, it computes the dot product of ALL vector pairs in the two inputs.

The **np.matmul()** method, as the name suggests, is meant for matrices. Thus, it computes the matrix multiplication of corresponding matrices in the two inputs.



# Run SQL in Jupyter To Analyze A Pandas DataFrame

Filter-Pandas.ipynb

```
df[df.city == "New Delhi"]
```

Pandas

DuckDB

Filter-SQL.ipynb

```
%%sql  
select * from df  
where city = 'New Delhi';
```

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

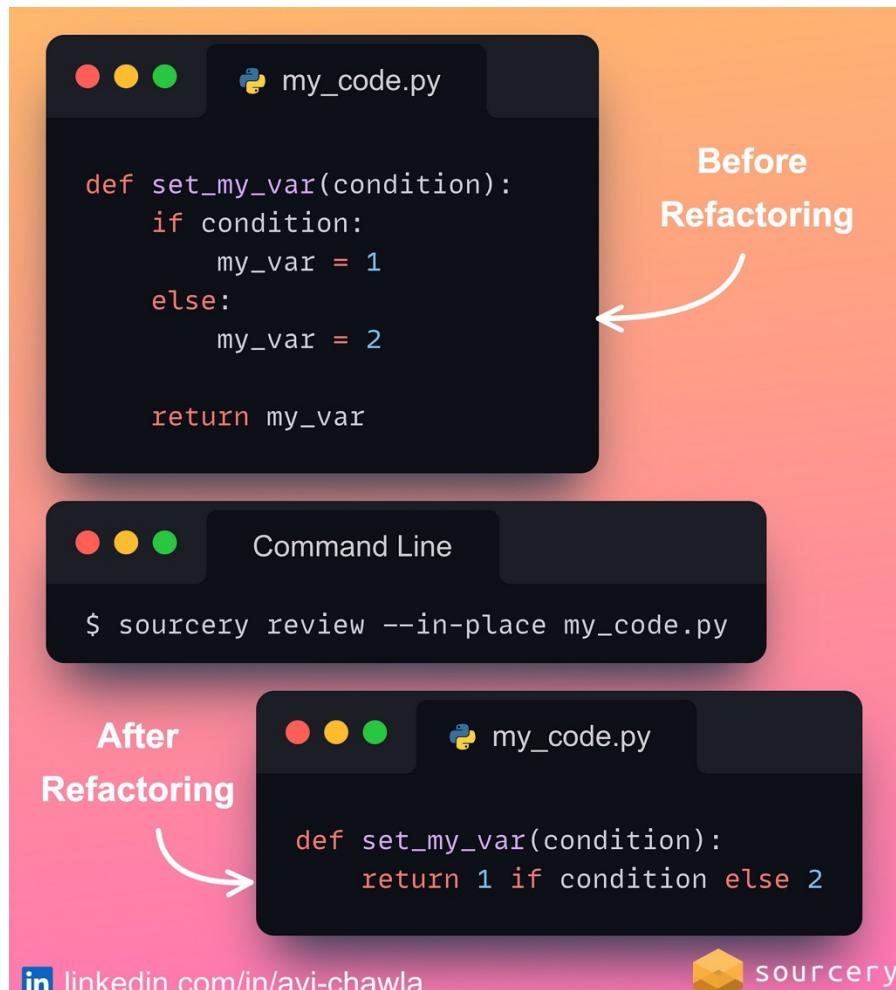
Pandas already provides a wide range of functionalities to analyze tabular data. Yet, there might be situations when one feels comfortable using SQL over Python.

Using DuckDB, you can analyze a Pandas DataFrame with SQL syntax in Jupyter, without any significant run-time difference.

Read the guide here to get started: [Docs](#).



# Automated Code Refactoring With Sourcery



[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla



Refactoring codebase is an important yet time-consuming task. Moreover, at times, one might unknowingly introduce errors during refactoring.

This takes additional time for testing and gets tedious with more refactoring, especially when the codebase is big.

Rather than following this approach, use [Sourcery](#). It's an automated refactoring tool that makes your code elegant, concise, and Pythonic in no time.

With Sourcery, you can refactor code in many ways. For instance, you can refactor scripts through the command line, as an IDE plugin in VS Code and PyCharm, etc.

Read my full blog on Sourcery here: [Medium](#).



# Post\_init\_: Add Attributes To A Dataclass Object Post Initialization

```
from dataclasses import dataclass

@dataclass
class StudentMarks:
    student_id:str
    marks:float

    def __post_init__(self):
        if self.marks>30:
            self.grade = "Pass"
        else:
            self.grade = "Fail"

Peter = StudentMarks("B20", 43)

print(Peter.grade) # Pass
```

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

After initializing a class object, we often create derived attributes from existing variables.

To do this in dataclasses, you can use the `__post_init__` method. As the name suggests, this method is invoked right after the `__init__` method.

This is useful if you need to perform additional setups on your dataclass instance.



# Simplify Your Functions With Partial Functions

Fix some parameters →

```
def quadratic(x, a, b, c):
    return a*(x**2) + b*x + c
```

```
partial_function.py
```

```
from functools import partial
quadratic_c1 = partial(quadratic, c=1)

quadratic_c1(x=1, a=4, b=5)
# Output: 10
```

[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

When your function takes many arguments, it can be a good idea to simplify it by using partial functions.

They let you create a new version of the function with some of the arguments fixed to specific values.

This can be useful for simplifying your code and making it more readable and concise. Moreover, it also helps you avoid repeating yourself while invoking functions.



# When You Should Not Use the `head()` Method In Pandas

`df.sort_values(by="Marks", ascending=False).head(3)`

	Name	Marks
1	Jane	100
2	Mark	97
0	Peter	95

**Ignores repeated values**

`df`

	Name	Marks
0	Peter	95
1	Jane	100
2	Mark	97
3	David	95

`df.nlargest(n=3, columns="Marks", keep="all")`

	Name	Marks
1	Jane	100
2	Mark	97
0	Peter	95
3	David	95

**Returns Duplicate values**

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

One often retrieves the top **k** rows of a sorted Pandas DataFrame by using `head()` method. However, there's a flaw in this approach.

If your data has repeated values, `head()` will not consider that and just return the first **k** rows.

If you want to consider repeated values, use `nlargest` (or `nsmallest`) instead. Here, you can specify the desired behavior for duplicate values using the `keep` parameter.



# DotMap: A Better Alternative to Python Dictionary

The image shows two terminal windows side-by-side. The left window, titled 'dotmap.py', contains Python code for creating a hierarchical data structure using the DotMap class. The right window, titled 'dotmap-print.py', shows the pretty printed output of this structure.

**dotmap.py:**

```
from dotmap import DotMap

students = DotMap()

students.john.id = "12A"
students.john.english = 45

students.mary.id = "12B"
students.mary.english = 49

students.dave.id = "12C"
students.dave.english = 34
```

**dotmap-print.py:**

```
people pprint()

{'dave': {'english': 34, 'id': '12C'},
 'john': {'english': 45, 'id': '12A'},
 'mary': {'english': 49, 'id': '12B'}}
```

**Add using dot notation** (with a curved arrow pointing from the text to the code)

**Pretty Print** (with a curved arrow pointing from the text to the code)

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Python dictionaries are great, but they have many limitations.

It is difficult to create dynamic hierarchical data. Also, they don't offer the widely adopted dot notation to access values.

Instead, use DotMap. It behaves like a Python dictionary but also addresses the above limitations.

What's more, it also has a built-in pretty print method to display it as a dict/JSON for debugging large objects.

Read more: [GitHub](#).



# Prevent Wild Imports With `__all__` in Python

The diagram illustrates the use of `__all__` to prevent wild imports. It shows two code snippets: `my_functions.py` and `module.py`.

**my\_functions.py:**

```
__all__ = ["func1", "func2"]

def func1():
    return "Function 1"

def func2():
    return "Function 2"

def func3():
    return "Function 3"
```

A callout points from the `__all__` assignment to the text "Specify `__all__`".

**module.py:**

```
from my_functions import *

>>> func1()
"Function 1"

>>> func2()
"Function 2"

>>> func3()
NameError: name 'func3' is not defined
```

A callout points from the `__all__` assignment in `my_functions.py` to the text "Only func1 and func2 imported".

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Wild imports (`from module import *`) are considered a bad programming practice. Yet, here's how you can prevent it if someone irresponsibly does that while using your code.

In your module, you can define the importable functions/classes/variables in `__all__`. As a result, whenever someone will do a wild import, Python will only import the symbols specified here.

This can be also useful to convey what symbols in your module are intended to be private.



# Three Lesser-known Tips For Reading a CSV File Using Pandas

The image shows three screenshots of a Mac terminal window with a dark theme. Each screenshot displays a different Pandas command for reading a CSV file:

- Top Screenshot:** Shows the command `pd.read_csv("data.csv", nrows = 10)`. To its right, the text "Read only first 10 rows" is displayed.
- Middle Screenshot:** Shows the command `pd.read_csv("data.csv", usecols = ["A", "C"])`. To its left, the text "Read specific columns" is displayed.
- Bottom Screenshot:** Shows two commands:
  - `pd.read_csv("data.csv", skiprows = 10)` with the annotation "Skip first 10 rows".
  - `pd.read_csv("data.csv", skiprows = [1, 5])` with the annotation "Skip 1st row and 5th row".

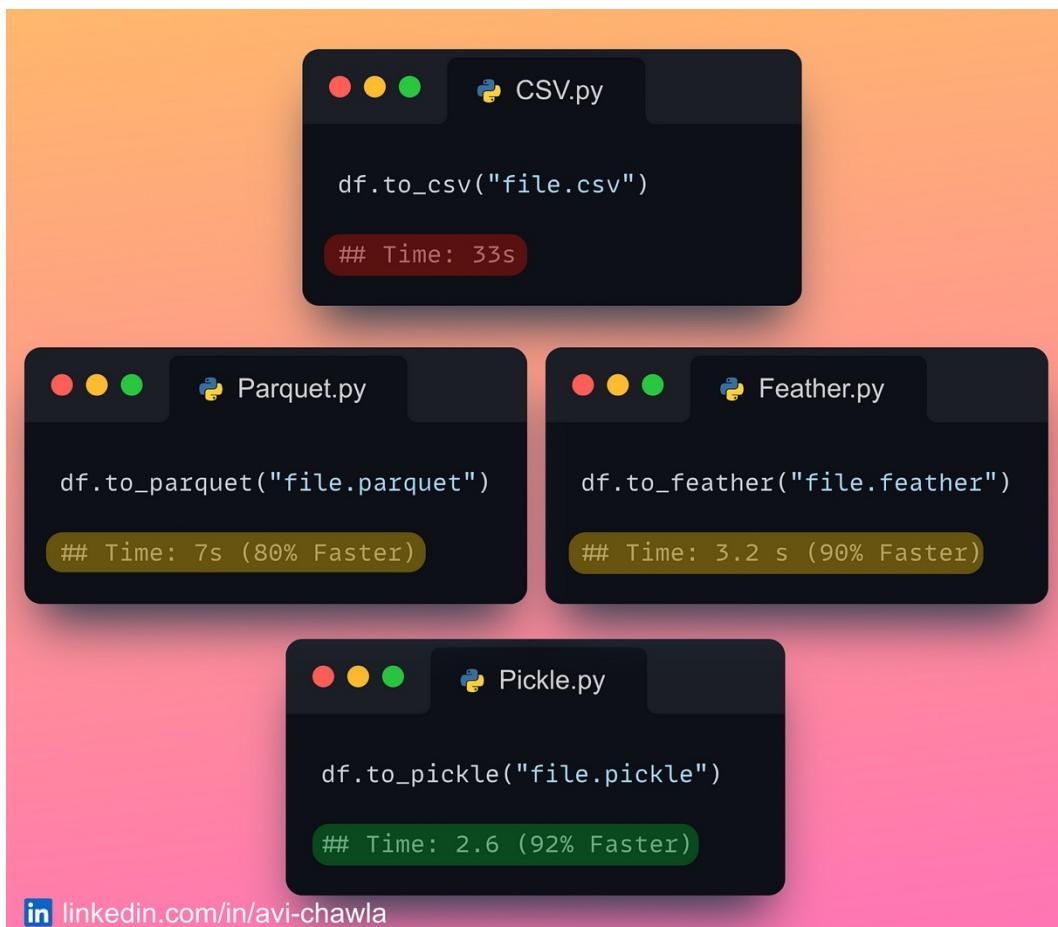
At the bottom left of the terminal window, there is a LinkedIn icon followed by the URL [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla).

Here are three extremely useful yet lesser-known tips for reading a CSV file with Pandas:

1. If you want to read only the first few rows of the file, specify the **nrows** parameter.
2. To load a few specific columns, specify the **usecols** parameter.
3. If you want to skip some rows while reading, pass the **skiprows** parameter.



# The Best File Format To Store A Pandas DataFrame



In the image above, you can find the run-time comparison of storing a Pandas DataFrame in various file formats.

Although CSVs are a widely adopted format, it is the slowest format in this list.

Thus, CSVs should be avoided unless you want to open the data outside Python (in Excel, for instance).

Read more in my blog: [Medium](#).



# Debugging Made Easy With PySnooper

The image shows two screenshots demonstrating PySnooper. The top screenshot is a code editor window titled 'py-snooper.py' containing Python code. A line of code is highlighted with a green background: `3 @pysnooper.snoop()`. An annotation 'Add Decorator' with a curved arrow points to this line. The bottom screenshot shows the terminal output of running the script: `$ python py-snooper.py`. The output shows variable values at each step of the function execution: 'Starting var:... a = 9', 'Starting var:... b = 5', 'New var:..... add = 14', 'New var:..... sub = 4', and 'Return value:.. (14, 4)'. An annotation 'Debugging Output' with a curved arrow points to the terminal output.

```
1 import pysnooper
2
3 @pysnooper.snoop()
4 def add_sub(a, b):
5
6     add = a+b
7     sub = a-b
8
9     return (add, sub)
10
11 add_sub(9, 5)
```

```
$ python py-snooper.py
Starting var:... a = 9
Starting var:... b = 5
New var:..... add = 14
New var:..... sub = 4
Return value:.. (14, 4)
```

[in/avi-chawla](https://linkedin.com/in/avi-chawla)

Rather than using many print statements to debug your python code, try PySnooper.

With just a single line of code, you can easily track the variables at each step of your code's execution.

Read more: [Repository](#).



# Lesser-Known Feature of the Merge Method in Pandas

The screenshot shows a Jupyter Notebook cell with the following code:

```
pd.merge(name_df, rewards_df,  
         on = "Cust_ID",  
         how = "outer",  
         indicator = True)
```

An arrow points from the `indicator = True` part of the code to a green column in the resulting DataFrame, which is labeled **Indicator Column**.

Cust_ID	Name	Rewards	_merge
1	Joe	NaN	left_only
2	Mark	50	both
3	Peter	20	both
4	NaN	70	right_only

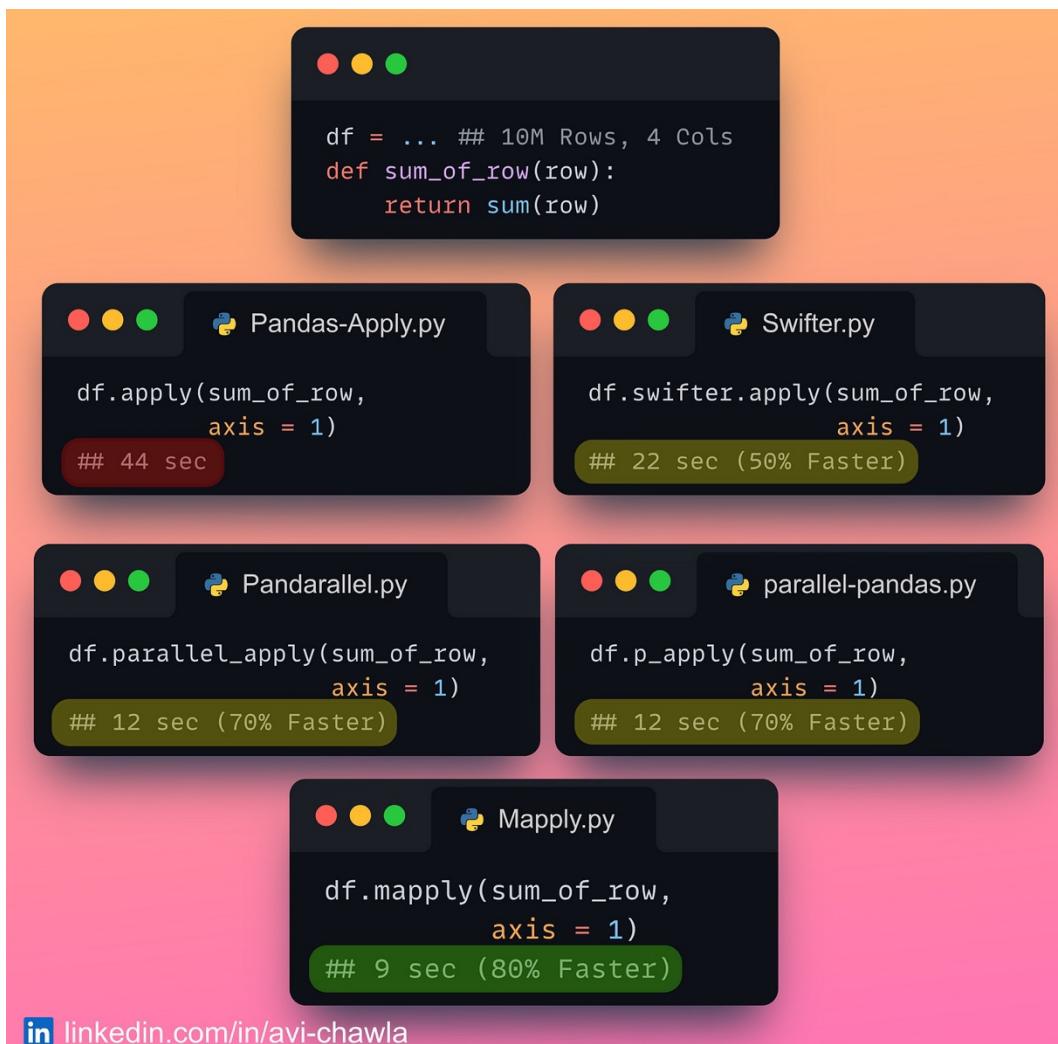
[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

While merging DataFrames in Pandas, keeping track of the source of each row in the output can be extremely useful.

You can do this using the `indicator` argument of the `merge()` method. As a result, it augments an additional column in the merged output, which tells the source of each row.



# The Best Way to Use Apply() in Pandas



The image above shows a run-time comparison of popular open-source libraries that provide parallelization support for Pandas.

You can find the links to these libraries [here](#). Also, if you know any other similar libraries built on top of Pandas, do post them in the comments or reply to this email.



# Deep Learning Network Debugging Made Easy

```
import tsensor

W = # Shape: (n_neurons, hidden_size)
b = # Shape: (n_neurons, 1)
X = # Shape: (n_batches, batch_size, hidden_size)

with tsensor.explain():
    for i in range(n_batches):
        batch = X[i,:,:]
        Y = torch.matmul(W, batch.T) + b
```

**Output**

$\text{batch} = \underline{X} [i,:,:]$

$\begin{array}{c} 764 \\ 10 \\ \hline \text{float32} \end{array}$      $\begin{array}{c} 10 \\ 20 \\ \hline \text{float32} \end{array}$

$\begin{array}{c} Y \\ 10 \\ \hline \text{float32} \end{array} = \text{torch.matmul}(\underline{W}, \underline{\text{batch.T}}) + \underline{b}$

$\begin{array}{c} 100 \\ \hline \text{float32} \end{array}$      $\begin{array}{c} 764 \\ 100 \\ \hline \text{float32} \end{array}$      $\begin{array}{c} 764 \\ 10 \\ \hline \text{float32} \end{array}$      $\begin{array}{c} 100 \\ 1 \\ \hline \text{float32} \end{array}$

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Aligning the shape of tensors (or vectors/matrices) in a network can be challenging at times.

As the network grows, it is common to lose track of dimensionalities in a complex expression.

Instead of explicitly printing tensor shapes to debug, use **TensorSensor**. It generates an elegant visualization for each statement executed within its block. This makes dimensionality tracking effortless and quick.

In case of errors, it augments default error messages with more helpful details. This further speeds up the debugging process.

Read more: [Documentation](#)



# Don't Print NumPy Arrays! Use Lovely-NumPy Instead.

The image shows two terminal windows side-by-side. The left window displays raw NumPy array output, while the right window shows the output of the `lo` function from the `lovely_numpy` library. A callout bubble points from the text "Only numbers" to the right window's output. Another callout bubble points from the text "Summary of Array" to the right window's output. A LinkedIn link is at the bottom.

Only numbers

```
>>> array = np.random.rand(...)  
>>> array  
tensor([[0.59, 0.03, ..., 0.44, 0.41],  
       [0.60, 0.72, ..., 0.92, 0.61],  
       ...,  
       [0.57, 0.98, ..., 0.01, 0.91],  
       [0.00, 0.53, ..., 0.54, 0.54]])
```

Summary of Array

```
from lovely_numpy import lo  
  
>>> array = np.random.rand(...)  
>>> lo(array)  
array[10, 20] n=200  
x ∈ [0.0, 1.0] μ=0.51 σ=0.3  
  
>>> array = np.zeros(...)  
>>> lo(array)  
array[10] all_zeros  
  
>>> array = # With NaN and Inf  
>>> lo(array)  
array[10, 20] n=200  
x ∈ [0.0, 1.0] μ=0.51 σ=0.3 +Inf! NaN!
```

linkedin.com/in/avi-chawla

We often print raw numpy arrays during debugging. But this approach is not very useful. This is because printing does not convey much information about the data it holds, especially when the array is large.

Instead, use **lovely-numpy**. Rather than viewing raw arrays, it prints a summary of the array. This includes its shape, distribution, mean, standard deviation, etc.

It also shows if the numpy array has NaNs and Inf values, whether it is filled with zeros, and many more.

P.S. If you work with tensors, then you can use **lovely-tensors**.

Read more: [Documentation](#).



# Performance Comparison of Python 3.11 and Python 3.10

The image shows a Mac desktop with four terminal windows side-by-side, each displaying Python code and its execution time.

- Top Left:** Terminal window titled "fib.py". It contains a Fibonacci function definition. The code is:

```
def fib(N):
    """
    Function to find the
    Nth Fibonacci number.

    fib(N) = fib(N-1) + fib(N-2)
    """
    ...
```
- Top Right:** Terminal window titled "calc\_pi.py". It contains a pi approximation function definition. The code is:

```
def pi_approx(n_terms):
    """
    Function to find the
    approximate value of pi.

    pi = 4*(1 - 1/3 + 1/5 - 1/7...)
    """
    ...
```
- Bottom Left:** Terminal window titled "python3.10.py". It shows three command-line executions of the code from the top windows, with their times:

```
>>> fib(30)
# Time: 260ms

>>> fib(40)
# Time: 32s

>>> pi_approx(10**6)
# Time: 144ms
```
- Bottom Right:** Terminal window titled "python3.11.py". It shows the same three command-line executions, with significantly faster times:

```
>>> fib(30)
# Time: 97ms (62% Faster)

>>> fib(40)
# Time: 12s (62% Faster)

>>> pi_approx(10**6)
# Time: 65ms (55% Faster)
```

Arrows point from the "Time:" labels in the bottom-left window to the corresponding "Time:" labels in the bottom-right window.

[in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Python 3.11 was released recently, and as per the official release, it is expected to be 10-60% faster than Python 3.10.

I ran a few basic benchmarking experiments to verify the performance boost. Indeed, Python 3.11 is much faster.

Although one might be tempted to upgrade asap, there are a few things you should know. Read more [here](#).



# View Documentation in Jupyter Notebook



[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

While working in Jupyter, it is common to forget the parameters of a function and visit the official docs (or Stackoverflow). However, you can view the documentation in the notebook itself.

Pressing **Shift-Tab** opens the documentation panel. This is extremely useful and saves time as one does not have to open the official docs every single time.

This feature also works for your custom functions.

View a video version of this post on LinkedIn: [Post Link](#).



# A No-code Tool To Understand Your Data Quickly

The screenshot shows a Jupyter Notebook cell containing the following Python code:

```
from pandas_profiling import ProfileReport

profile = ProfileReport(iris_data,
                       title="Pandas Profiling Report")

profile.to_widgets()
```

Below the code is a screenshot of the generated Pandas Profiling Report. The report has a tabbed interface with "Overview" selected. Under "Alerts (7)", it lists several correlations and distributions:

- Dataset has 1 (0.7%) duplicate rows
- sepal length (cm) is highly correlated with sepal width (cm) and 3.other.fields
- petal length (cm) is highly correlated with sepal length (cm) and 3.other.fields
- petal width (cm) is highly correlated with sepal length (cm) and 3.other.fields
- target is highly correlated with sepal length (cm) and 3.other.fields
- sepal width (cm) is highly correlated with sepal length (cm) and 3.other.fields
- target is uniformly distributed

On the right side of the alerts, there are categories: Duplicates, High correlation, and Uniform. The "Uniform" category is highlighted in red.

Report generated by YData.

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

The preliminary steps of any typical EDA task are often the same. Yet, across projects, we tend to write the same code to carry out these tasks. This gets repetitive and time-consuming.

Instead, use **pandas-profiling**. It automatically generates a standardized report for data understanding in no time. Its intuitive UI makes this effortless and quick.

The report includes the dimension of the data, missing value stats, and column data types. What's more, it also shows the data distribution, the interaction and correlation between variables, etc.

Lastly, the report also includes alerts, which can be extremely useful during analysis/modeling.

Read more: [Documentation](#).



# Why 256 is 256 But 257 is not 257?

The screenshot shows an IPython terminal window with a dark theme. The terminal has three colored tabs at the top: red, yellow, and green. The green tab is active and displays the IPython logo. The terminal window contains the following Python code:

```
>>> a = 256
>>> b = 256

>>> a is b
True

>>> a = 257
>>> b = 257

>>> a is b
False

>>> a, b = 257, 257

>>> a is b
True
```

Below the terminal window, there is a blue LinkedIn icon followed by the URL [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla).

Comparing python objects can be tricky at times. Can you figure out what is going on in the above code example? Answer below:

---

When we run Python, it pre-loads a global list of integers in the range [-5, 256]. Every time an integer is referenced in this range, Python does not create a new object. Instead, it uses the cached version.

This is done for optimization purposes. It was considered that these numbers are used a lot by programmers. Therefore, it would make sense to have them ready at startup.



However, referencing any integer beyond 256 (or before -5) will create a new object every time.

In the last example, when a and b are set to 257 in the same line, the Python interpreter creates a new object. Then it references the second variable with the same object.

Share this post on LinkedIn: [Post Link](#).

The below image should give you a better understanding:



[in](https://linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla



# Make a Class Object Behave Like a Function

The diagram illustrates how to make a class object callable. On the left, a code editor shows the definition of a `Quadratic` class with an `__init__` method and a `__call__` method. A callout bubble points to the `__call__` method with the text "define \_\_call\_\_ method". On the right, a terminal window demonstrates creating an instance of the class and calling it like a function, with output showing the quadratic formula results for x=1 and x=2. Another callout bubble points to the terminal window with the text "class object behaves like function".

```
class Quadratic:
    def __init__(self, a, b, c):
        self.a = a
        self.b = b
        self.c = c

    def __call__(self, x):
        return (self.a * x**2) +
               (self.b * x) +
               self.c
```

```
f = Quadratic(1, 2, 3)
print(f(1)) # Output: 6
print(f(2)) # Output: 11
print(callable(f)) # Output: True
```

class object behaves like function

define \_\_call\_\_ method

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

If you want to make a class object callable, i.e., behave like a function, you can do so by defining the `__call__` method.

This method allows you to define the behavior of the object when it is invoked like a function.



[avichawla.substack.com](https://avichawla.substack.com)

This can have many advantages. For instance, it allows us to implement objects that can be used in a flexible and intuitive way. What's more, the familiar function-call syntax, at times, can make your code more readable.

Lastly, it allows you to use a class object in contexts where a callable is expected. Using a class as a decorator, for instance.



# Lesser-known feature of Pickle Files

The image shows two terminal windows on a Mac OS X desktop. The top window, titled 'dump.py', contains Python code that writes three variables (a, b, c) to a file named 'data.pkl'. The bottom window, titled 'load.py', contains Python code that reads these variables back from the file. A white arrow points from the text 'Store 3 Variables' to the 'with' block in the dump.py code. Another white arrow points from the text 'Load Only First 2 Variables' to the 'with' block in the load.py code.

```
import pickle

a, b, c = 1, 2, 3

with open("data.pkl", "wb") as f:
    pickle.dump(a, f)
    pickle.dump(b, f)
    pickle.dump(c, f)
```

```
import pickle

with open("data.pkl", "rb") as f:
    a = pickle.load(f)
    b = pickle.load(f)

print(f"{a = } {b = }")
## a = 1 b = 2
```

**Store 3 Variables**

**Load Only First 2 → Variables**

Pickles are widely used to dump data objects to disk. But folks often dump just a single object into a pickle file. Moreover, one creates multiple pickles to store multiple objects.

However, did you know that you can store as many objects as you want within a single pickle file? What's more, when reloading, it is not necessary to load all the objects.

Just make sure to dump the objects within the same context manager (using **with**).



Of course, one solution is to store the objects together as a tuple. But while reloading, the entire tuple will be loaded. This may not be desired in some cases.



# Dot Plot: A Potential Alternative to Bar Plot



Bar plots are extremely useful for visualizing categorical variables against a continuous value. But when you have many categories to depict, they can get too dense to interpret.

In a bar plot with many bars, we're often not paying attention to the individual bar lengths. Instead, we mostly consider the individual endpoints that denote the total value.

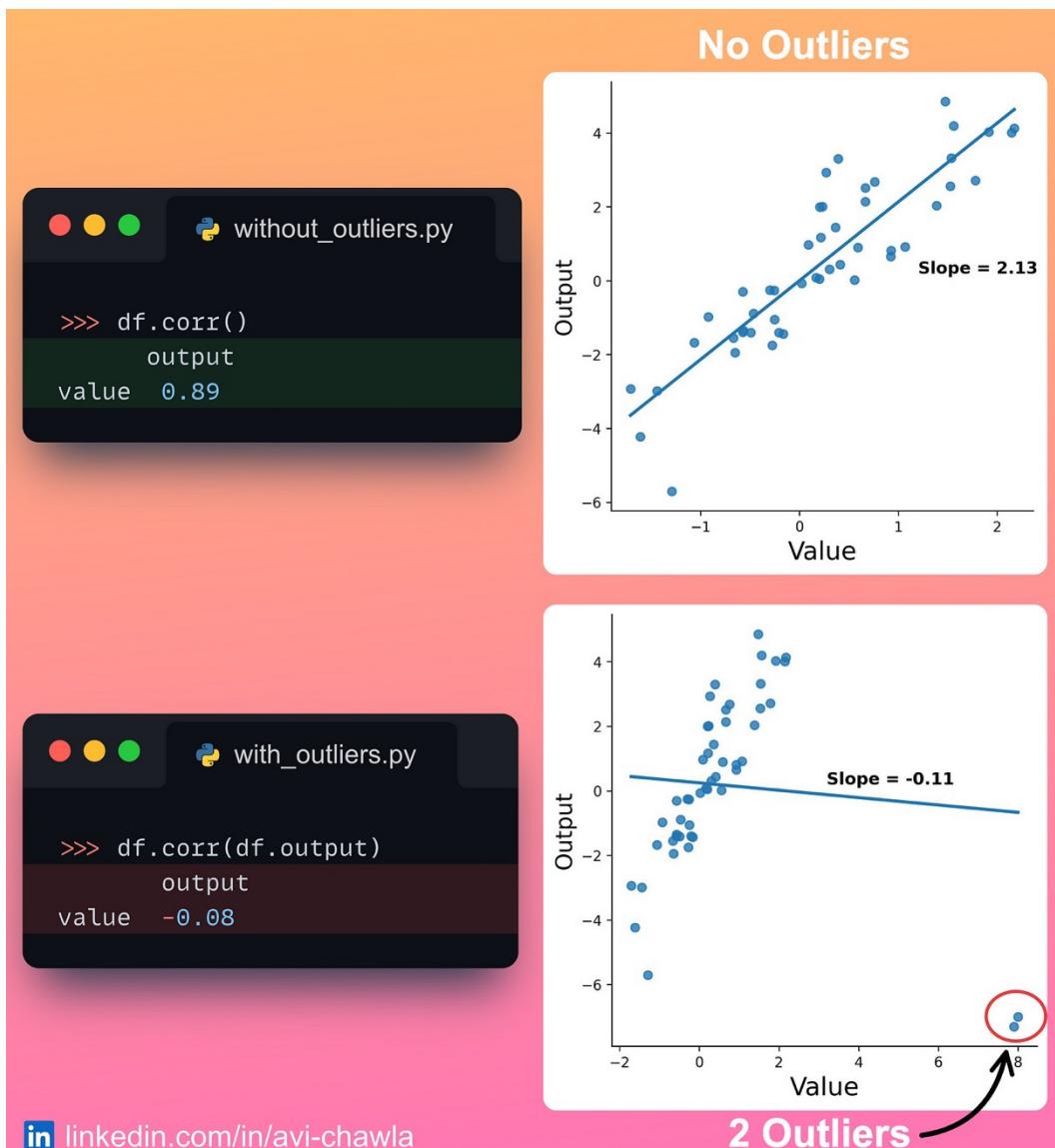
A Dot plot can be a better choice in such cases. They are like scatter plots but with one categorical and one continuous axis.

Compared to a bar plot, they are less cluttered and offer better comprehension. This is especially true in cases where we have many categories and/or multiple categorical columns to depict in a plot.

Read more: [Documentation](#).



# Why Correlation (and Other Statistics) Can Be Misleading.



Correlation is often used to determine the association between two continuous variables. But it has a major flaw that often gets unnoticed.

Folks often draw conclusions using a correlation matrix without even looking at the data. However, the obtained statistics could be heavily driven by outliers or other artifacts.

This is demonstrated in the plots above. The addition of just two outliers changed the correlation and the regression line drastically.

Thus, looking at the data and understanding its underlying characteristics can save from drawing wrong conclusions. Statistics are important, but they can be highly misleading at times.



# Supercharge value\_counts() Method in Pandas With Sidetable

```
import sidetable
df.stb.freq(["City"], style=True)
```

	City	count	percent	cumulative_count	cumulative_percent
0	West Jamesview	120	12.00%	120	12.00%
1	Aliciafort	113	11.30%	233	23.30%
2	Ricardomouth	106	10.60%	339	33.90%
3	New Cindychester	106	10.60%	445	44.50%
4	Whiteside	104	10.40%	549	54.90%
5	Kristaburgh	97	9.70%	646	64.60%
6	Wardfort	96	9.60%	742	74.20%
7	New Russellton	93	9.30%	835	83.50%
8	Whitakerbury	87	8.70%	922	92.20%
9	North Melissafurt	78	7.80%	1,000	100.00%

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

The **value\_counts()** method is commonly used to analyze categorical columns, but it has many limitations.

For instance, if one wants to view the percentage, cumulative count, etc., in one place, things do get a bit tedious. This requires more code and is time-consuming.

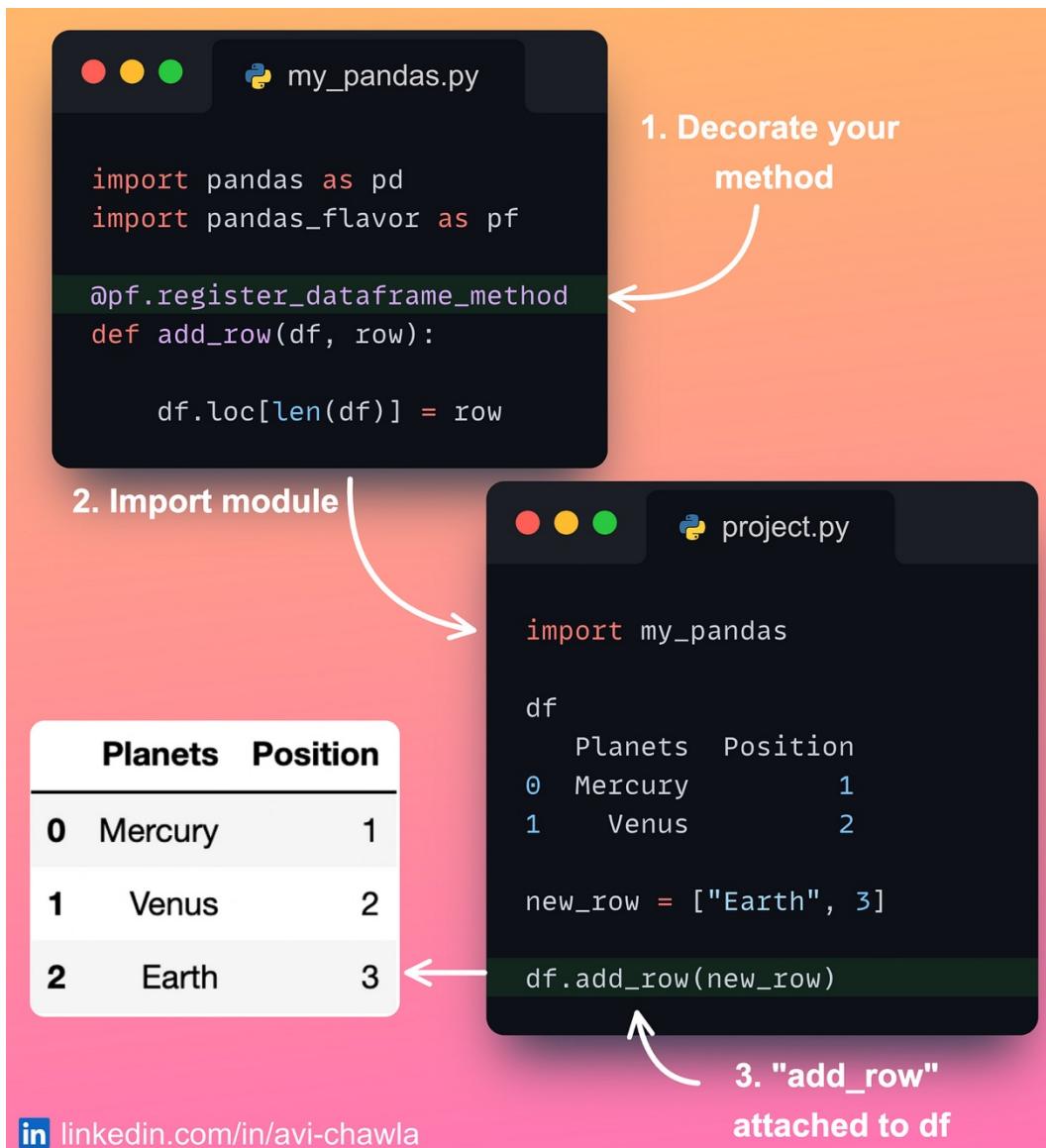
Instead, use **sidetable**. Consider it as a supercharged version of **value\_counts()**. As shown below, the **freq()** method from sidetable provides a more useful summary than **value\_counts()**.

Additionally, sidetable can aggregate multiple columns too. You can also provide threshold points to merge data into a single bucket. What's more, it can print missing data stats, pretty print values, etc.

Read more: [GitHub](#).



# Write Your Own Flavor Of Pandas



If you want to attach a custom functionality to a Pandas DataFrame (or series) object, use "pandas-flavor".

Its decorators allow you to add methods directly to the Pandas' object.

This is especially useful if you are building an open-source project involving Pandas. After installing your library, others can access your library's methods using the `dataframe` object.

P.S. This is how we see `df.progress_apply()` from `tqdm`, `df.parallel_apply()` from `Pandarallel`, and many more.

Read more: [Documentation](#).



# CodeSquire: The AI Coding Assistant You Should Use Over GitHub Copilot

The screenshot shows the CodeSquire.ai interface. On the left, there's a code editor window with Python code for a titanic dataset. A red arrow points from the text "Output" at the bottom right towards the code editor. To the right of the code editor, the text "Write comment" is displayed above a downward-pointing arrow. The background of the interface is orange and pink.

```
from catboost.datasets import titanic
import numpy as np
import pandas as pd
from catboost import CatBoostClassifier

train_df, test_df = titanic()

train_df.head()

# one hot encode all the categorical vars in train_df and test_df
train_df = pd.get_dummies(train_df, columns=['Sex', 'Embarked'])
test_df = pd.get_dummies(test_df, columns=['Sex', 'Embarked'])
```

Output

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Coding Assistants like GitHub Copilot are revolutionary as they offer many advantages. Yet, Copilot has limited utility for data professionals. This is because it's incompatible with web-based IDEs (Jupyter/Colab).

Moreover, in data science, the subsequent exploratory steps are determined by previous outputs. But Copilot does not consider that (and even markdown cells) to drive its code suggestions.

[\*\*CodeSquire\*\*](#) is an incredible AI coding assistant that addresses the limitations of Copilot. The good thing is that it has been designed specifically for data scientists, engineers, and analysts.

Besides seamless code generation, it can generate SQL queries from text and explain code. You can leverage AI-driven code generation by simply installing a browser extension.

Read more: [\*\*CodeSquire\*\*](#).

Watch a video version of this post on LinkedIn: [Post Link](#).



# Vectorization Does Not Always Guarantee Better Performance

The screenshot shows a Mac desktop with three main components:

- Terminal Window 1 (Vectorized.py):** Contains the command `df.Name.str.split()` and its execution time `## 1.7 s`.
- Terminal Window 2 (Non-vectorized.py):** Contains the Python code for a function `name_split` that splits strings and its execution time `## 862 ms`.
- DataFrame (df):** A table with columns `df` and `Name`, containing 5 rows of data.

df	Name
0	Beth Alvarez
1	Deborah Watkins
2	Jeffrey Compton
3	Alan Wolfe
4	Kathryn Gordon

[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Vectorization is well-adopted for improving run-time performance. In a nutshell, it lets you operate data in batches instead of processing a single value at a time.

Although vectorization is extremely effective, you should know that it does not always guarantee performance gains. Moreover, vectorization is also associated with memory overheads.

As demonstrated above, the non-vectorized code provides better performance than the vectorized version.

P.S. `apply()` is also a for-loop.

Further reading: [Here](#).



# In Defense of Match-case Statements in Python

**if-else.py**

```
def make_point(point):
    if isinstance(point, (tuple, list)):
        if len(point) == 2:
            x, y = point
            return Point3D(x, y, 0)

        elif len(point) == 3:
            x, y, z = point
            return Point3D(x, y, z)

        else:
            raise TypeError("Unsupported")
    else:
        raise TypeError("Unsupported")
```

**Check type**

**Check length**

**Explicit unpacking**

**No type checks**

**No length checks**

**No unpacking**

**match-case.py**

```
def make_point(point):
    match point:
        case (x, y):
            return Point3D(x, y, 0)

        case (x, y, z):
            return Point3D(x, y, z)

        case _: ## Default
            raise TypeError("Unsupported")

    >>> make_point((1, 2))
    Point3D(x=1, y=2, z=0)

    >>> make_point([1, 2, 3])
    Point3D(x=1, y=2, z=0)

    >>> make_point((1, 2, 3, 4))
    TypeError: Unsupported
```

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)



I recently came across a post on **match-case** in Python. In a gist, starting Python 3.10, you can use **match-case** statements to mimic the behavior of **if-else**.

Many responses on that post suggested that **if-else** offers higher elegance and readability. Here's an example in defense of **match-case**.

While if-else is traditionally accepted, it also comes with many downsides. For instance, many-a-times, one has to write complex chains of nested if-else statements. This includes multiple calls to **len()**, **isinstance()** methods, etc.

Furthermore, with **if-else**, one has to explicitly destructure the data to extract values. This makes your code inelegant and messy.

Match-case, on the other hand, offers Structural Pattern Matching which makes this simple and concise. In the example above, match-case automatically handles type-matching, length check, and variable unpacking.

Read more here: [Python Docs](#).



# Enrich Your Notebook With Interactive Controls

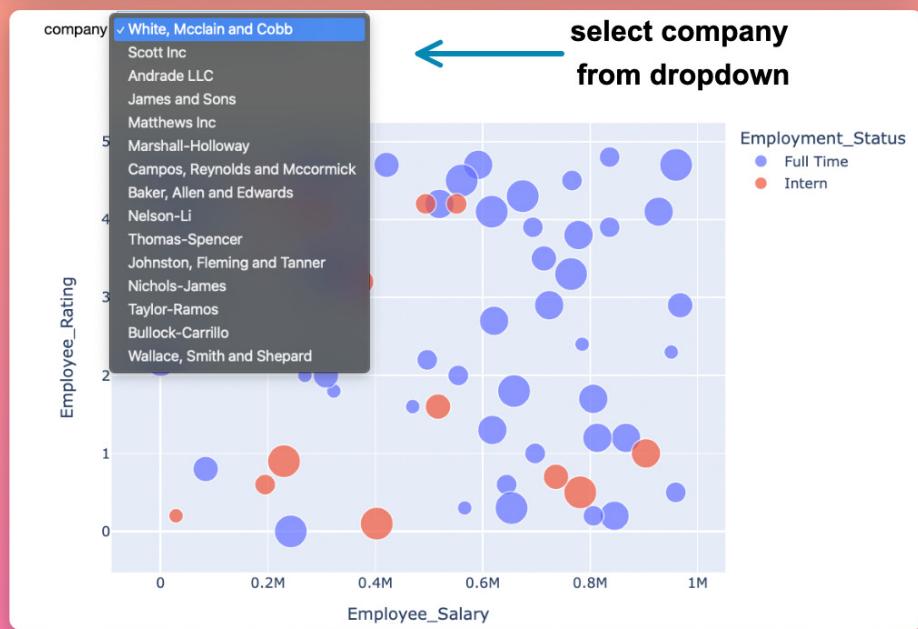
```
Add Drop-down

from ipywidgets import interact

@interact
def plot_company(company = list(df.Company.unique())):

    ## filter the DataFrame
    df_filtered = df[df.Company == company]

    ## Create the plot on df_filtered
    df_filtered.plot(...)
```



[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

While using Jupyter, we often re-run the same cell repeatedly after changing the input slightly. This is time-consuming and also makes your data exploration tasks tedious and unorganized.



Instead, pivot towards building interactive controls in your notebook. This allows you to alter the inputs without needing to rewrite and re-run your code.

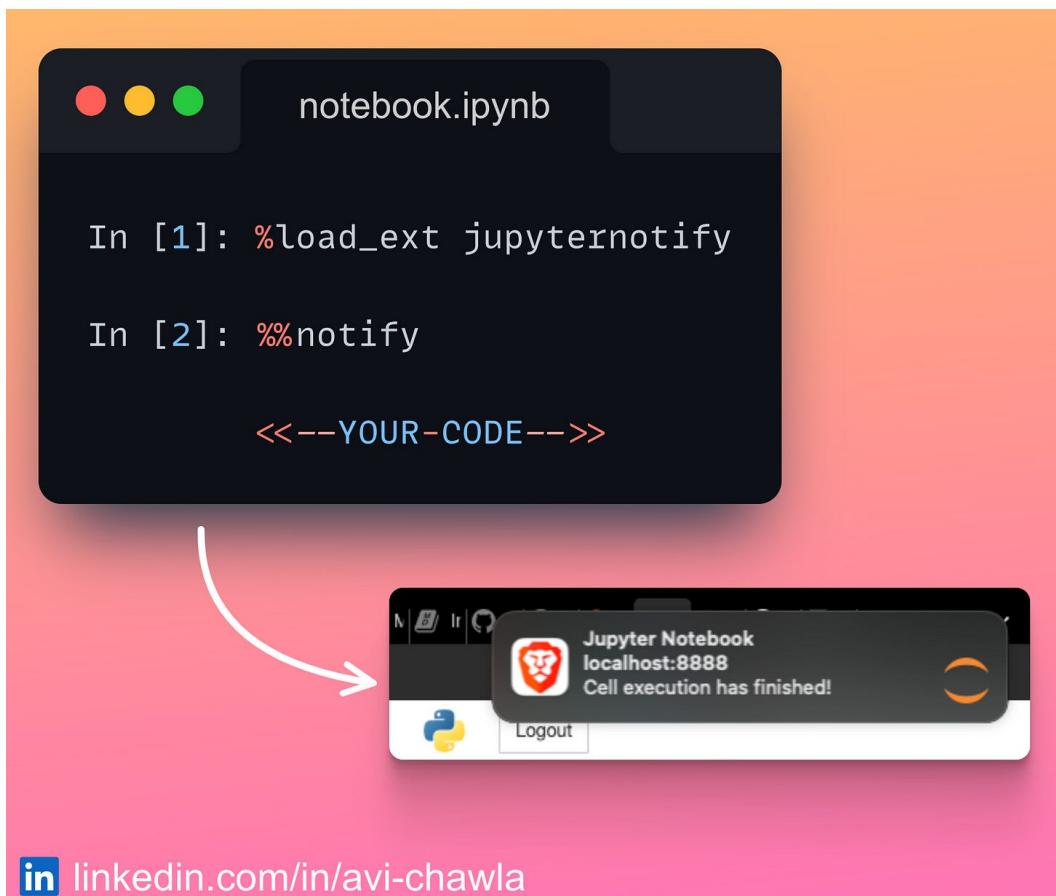
In Jupyter, you can do this using the **IPywidgets** module. Embedding interactive controls is as simple as using a decorator.

As a result, it provides you with interactive controls such as dropdowns and sliders. This saves you from tons of repetitive coding and makes your notebook organized.

Watch a video version of this post on LinkedIn: [Post Link](#).



# Get Notified When Jupyter Cell Has Executed



[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

After running some code in a Jupyter cell, we often navigate away to do some other work in the meantime.

Here, one has to repeatedly get back to the Jupyter tab to check whether the cell has been executed or not.

To avoid this, you can use the `%%notify` magic command from the `jupyternotify` extension. As the name suggests, it notifies the user upon completion (both successful and unsuccessful) of a jupyter cell via a browser notification. Clicking on the notification takes you back to the jupyter tab.

Read more: [GitHub](#).



# Data Analysis Using No-Code Pandas In Jupyter

The screenshot shows a Jupyter notebook cell with the following code:

```
In [1]: 1 import mitosheet
2 mitosheet.sheet.analysis_to_replay="id-ymyxvhaoes"
executed in 4.97s, finished 14:48:34 2022-11-22
```

Below the code, a Mito spreadsheet interface is displayed. The data is from a dataset named "employee\_dataset". The columns are:

	Name	Company_Nam	Employee_City	Employee_Sal	Employment_St	Employee_Rat
	str	str	str	float	str	float
99	Christopher Jones	Matthews Inc	Aliciafort	5,187.04	Full Time	1.50
96	Mitchell Hill	Baker, Allen and Edw	Aliciafort	4,078.71	Full Time	1.40
44	Dawn Bailey	White, McClain and C	Aliciafort	11,379.39	Full Time	4.50
80	Donald Bowman	Scott Inc	Aliciafort	4,292.43	Full Time	1.30
48	Kelly Liu	Matthews Inc	Aliciafort	4,413.51	Intern	0.90
20	David Mills	Johnston, Fleming an	Aliciafort	6,917.60	Intern	1.90
75	Vanessa Lamb	Taylor-Ramos	Aliciafort	8,391.54	Full Time	2.70
67	Douglas Kennedy	Andrade LLC	Aliciafort	2,815.63	Full Time	0.80
54	Jeffrey Gonzalez	Taylor-Ramos	Aliciafort	10,401.33	Full Time	4.60
37	Emily Weber	Matthews Inc	Kristaburgh	7,676.39	Intern	2.30
45	Zachary Ellison	James and Sons	Kristaburgh	7,194.54	Full Time	2.60
50	Gina Acosta	Nichols-James	Kristaburgh	7,239.98	Full Time	2.10
22	Jason Reyes	Matthews Inc	Kristaburgh	6,760.68	Full Time	2.80
53	James Wright	Nelson-Li	Kristaburgh	3,980.27	Intern	1.00

At the bottom of the interface, there are buttons for "+ employee\_dataset" and "graph0". To the right, it says "(100 rows, 6 cols)".

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

The Pandas API provides a wide range of functionalities to analyze tabular datasets.

Yet, across projects, we often use the same methods over and over to analyze our data. This quickly gets repetitive and time-consuming.

To avoid this, use Mito. It's an incredible tool that allows you to analyze your data within a spreadsheet interface in Jupyter, without writing any code.

The coolest thing about Mito is that each edit in the spreadsheet automatically generates an equivalent Python code. This makes it extremely convenient to reproduce the analysis later.

Read more: [Documentation](#).



# Using Dictionaries In Place of If-conditions

```
if_else.py
number = int(input())

if number == 1:
    func1()

elif number == 2:
    func2()

else:
    func3()
```

```
dict.py
number = int(input())

func_map = {1:func1,
            2:func2}

func_map.get(number, func3)()
```

replace with dictionary

Default

[in](https://linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Dictionaries are mainly used as a data structure in Python for maintaining key-value pairs.

However, there's another special use case that dictionaries can handle. This is — Eliminating IF conditions from your code.

Consider the code snippet above. Here, corresponding to an input value, we invoke a specific function. The traditional way requires you to hard-code every case.



[avichawla.substack.com](https://avichawla.substack.com)

But with a dictionary, you can directly retrieve the corresponding function by providing it with the key. This makes your code concise and elegant.



# Clear Cell Output In Jupyter Notebook During Run-time

The screenshot shows a Jupyter Notebook cell with the following Python code:

```
import time
from IPython.display import clear_output

for i in range(100):

    ## Wait for the next
    ## output before clearing
    clear_output(wait=True)

    print(f'Output Number {i+1}')
    time.sleep(1)
```

The cell has been executed, and the output panel shows:

In [6]:

```
1 for i in range(100):
2
3     ## Wait for the next
4     ## output before clearing
5     clear_output(wait=True)
6
7     print(f'Output Number {i+1}')
8     time.sleep(1)
```

executed in 1m 40.6s, finished 15:55:44 2022-11-19

Output Number 100 ← Only Last Output

[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

While using Jupyter, we often print many details to track the code's progress.

However, it gets frustrating when the output panel has accumulated a bunch of details, but we are only interested in the most recent output. Moreover, scrolling to the bottom of the output each time can be annoying too.

To clear the output of the cell, you can use the `clear_output` method from the `IPython` package. When invoked, it will remove the current output of the cell, after which you can print the latest details.



# A Hidden Feature of Describe Method In Pandas

The diagram illustrates the behavior of the `describe()` method in Pandas. It shows two Jupyter Notebook cells and their corresponding outputs.

**Only Numerical Columns:**

```
>>> df
      col1  col2  col3  col4
0      1      2      A      D
1      3      4      B      E
2      5      6      A      E

>>> df.describe()
```

	col1	col2
count	3.0	3.0
mean	3.0	4.0
std	2.0	2.0
min	1.0	2.0
25%	2.0	3.0
50%	3.0	4.0
75%	4.0	5.0
max	5.0	6.0

**All Columns:**

```
>>> df.describe(include = "all")
```

	col1	col2	col3	col4
count	3.0	3.0	3	3
unique	NaN	NaN	2	2
top	NaN	NaN	A	E
freq	NaN	NaN	2	2
mean	3.0	4.0	NaN	NaN
std	2.0	2.0	NaN	NaN
min	1.0	2.0	NaN	NaN
25%	2.0	3.0	NaN	NaN
50%	3.0	4.0	NaN	NaN
75%	4.0	5.0	NaN	NaN
max	5.0	6.0	NaN	NaN

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

The **`describe()`** method in Pandas is commonly used to print descriptive statistics about the data.

But have you ever noticed that its output is always limited to numerical columns? Of course, details like mean, median, std. dev., etc. hold no meaning for non-numeric columns, so the results make total sense.

However, **`describe()`** can also provide a quick summary of non-numeric columns. You can do this by specifying `include="all."` As a result, it will return the number of unique elements, the top element with its frequency.

Read more: [Documentation](#).



# Use Slotted Class To Improve Your Python Code

The image shows two code editors side-by-side. The left editor, titled 'Without\_slots.py', contains the following code:

```
class Person:  
    def __init__(self, name, age):  
        self.Name = name  
        self.Age = age  
  
person = Person('Mike', 22)  
  
person.name = 'Peter'  
## No Error
```

A callout bubble from the right side points to the line 'person.name = 'Peter'' with the text: "'Name' mistakenly written as 'name' raises no error'.

The right editor, titled 'With\_slots.py', contains the following code:

```
class Person:  
    __slots__ = ['Name', 'Age']  
  
    def __init__(self, name, age):  
        self.Name = name  
        self.Age = age  
  
person = Person('Mike', 22)  
  
person.name = 'Peter'  
## AttributeError: 'Person' object  
## has no attribute 'name'
```

A callout bubble from the left side points to the line 'person.name = 'Peter'' with the text: 'defining \_\_slots\_\_ raises AttributeError'.

At the bottom left of the image is a LinkedIn link: [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla).

If you want to fix the attributes a class can hold, consider defining it as a slotted class.

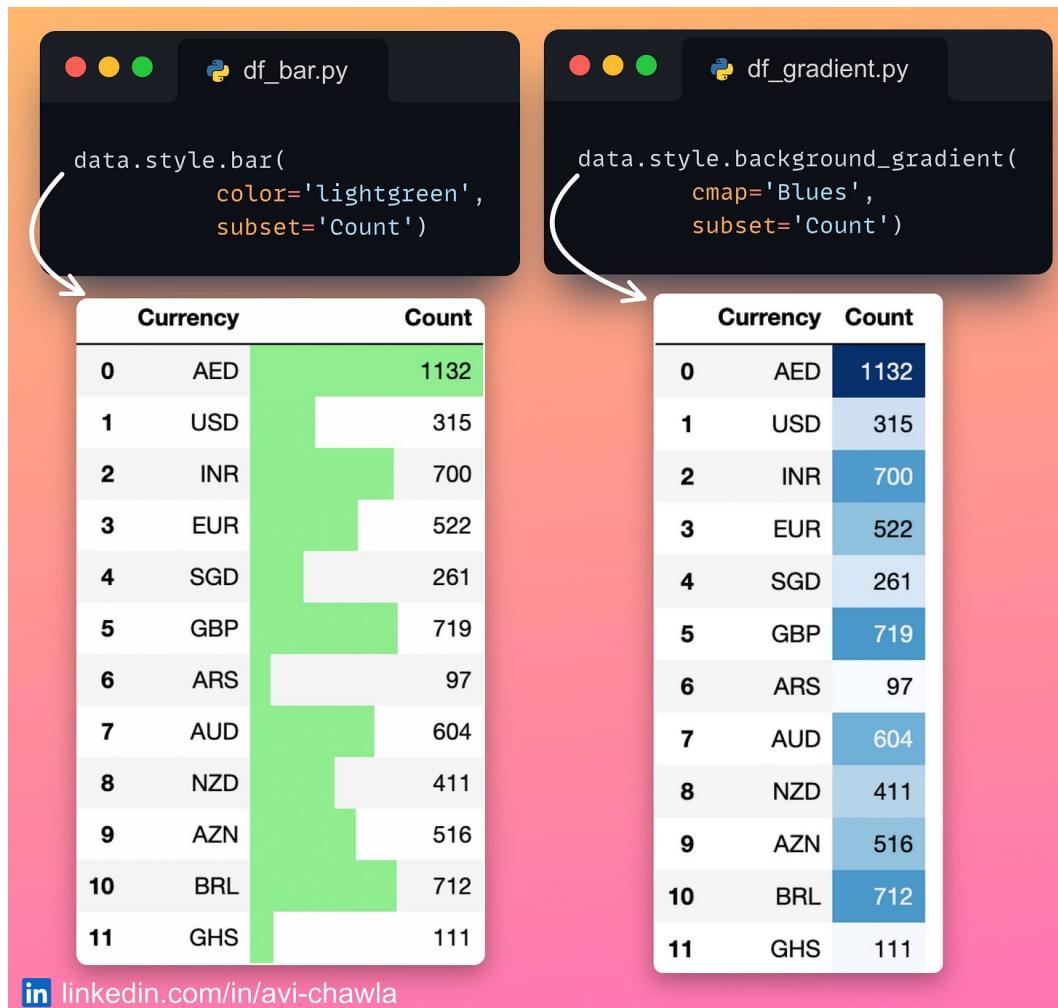
While defining classes, `__slots__` allows you to explicitly specify the class attributes. This means you cannot randomly add new attributes to a slotted class object. This offers many advantages.

For instance, slotted classes are memory efficient and they provide faster access to class attributes. What's more, it also helps you avoid common typos. This, at times, can be a costly mistake that can go unnoticed.

Read more: [StackOverflow](#).



# Stop Analysing Raw Tables. Use Styling Instead!



Jupyter is a web-based IDE. Thus, whenever you print/display a DataFrame in Jupyter, it is rendered using HTML and CSS.

This means you can style your output in many different ways.

To do so, use the Styling API of Pandas. Here, you can make many different modifications to a DataFrame's styler object (`df.style`). As a result, the DataFrame will be displayed with the specified styling.

Styling makes these tables visually appealing. Moreover, it allows for better comprehensibility of data than viewing raw tables.

Read more here: [Documentation](#).



# Explore CSV Data Right From The Terminal

The screenshot shows a terminal window with four distinct sections, each demonstrating a different command from the `csvkit` suite:

- Top Left:** Converts an Excel file to CSV.

```
$ in2csv data.xlsx > data.csv
```
- Top Right:** Displays the contents of the `data.csv` file.
- Middle Left:** Shows column names for the CSV file.
- Middle Right:** Provides column statistics for the `Marks` column.
- Bottom Left:** Runs a SQL query to select rows where `Marks > 90`.
- Bottom Right:** Shows the results of the SQL query.

[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

If you want to quickly explore some CSV data, you may not always need to run a Jupyter session.

Rather, with "`csvkit`", you can do it from the terminal itself. As the name suggests, it provides a bunch of command-line tools to facilitate data analysis tasks.

These include converting Excel to CSV, viewing column names, data statistics, and querying using SQL. Moreover, you can also perform popular Pandas functions such as sorting, merging, and slicing.

Read more: [Documentation](#).



# Generate Your Own Fake Data In Seconds

```
fake_data.py
```

```
from faker import Faker

fake = Faker()

>>> fake.name()
'Darrell Alexander'

>>> fake.email()
'ryanrichard@example.com'

>>> fake.address()
'205 Brown Point, West Melissaport, MN 93828'

>>> fake.company()
'Lam, Thomas and Cooper'

>>> fake.date_of_birth()
datetime.date(1973, 1, 21)

>>> fake.color_name()
'LightBlue'
```

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Usually, for executing/testing a pipeline, we need to provide it with some dummy data.

Although using Python's "**random**" library, one can generate random strings, floats, and integers. Yet, being random, it does not output any meaningful data such as people's names, city names, emails, etc.

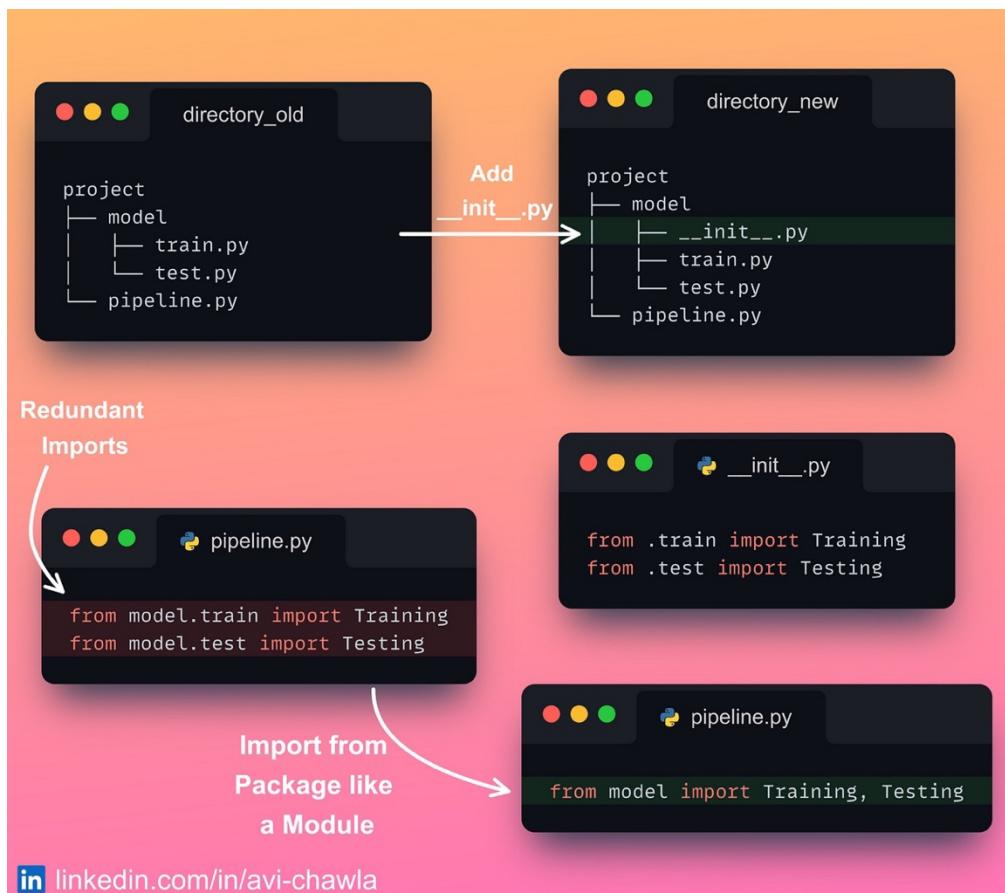
Here, looking for open-source datasets can get time-consuming. Moreover, it's possible that the dataset you find does not fit pretty well into your requirements.

The **Faker** module in Python is a perfect solution to this. Faker allows you to generate highly customized fake (yet meaningful) data quickly. What's more, you can also generate data specific to a demographic.

Read more here: [Documentation](#).



# Import Your Python Package as a Module



A python module is a single python file (**.py**). An organized collection of such python files is called a python package.

While developing large projects, it is a good practice to define an **\_\_init\_\_.py** file inside a package.

Consider **train.py** has a **Training** class and **test.py** has a **Testing** class.

Without **\_\_init\_\_.py**, one has to explicitly import them from specific python files. As a result, it is redundant to write the two import statements.

With **\_\_init\_\_.py**, you can group python files into a single importable module. In other words, it provides a mechanism to treat the whole package as a python module.

This saves you from writing redundant import statements and makes your code cleaner in the calling script.

Read more in this blog: [Blog Link](#).



# Specify Loops and Runs In `%%timeit`

The screenshot shows a Jupyter Notebook cell titled "notebook.ipynb". The cell contains the following code:

```
In [1]: %%timeit -n 1000 -r 4
    time.sleep(2)

## 2 s ± 800 µs per loop
## (mean ± std. dev. of 4 runs, 1000 loops each)
```

Annotations with arrows point to the parameters in the command:

- A double-headed arrow points between the text "number of" and "loops (1000)".
- A double-headed arrow points between the text "number of" and "runs (4)".

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

We commonly use the `%timeit` (or `%%timeit`) magic command to measure the execution time of our code.

Here, `timeit` limits the number of runs depending on how long the script takes to execute. This is why you get to see a different number of loops (and runs) across different pieces of code.

However, if you want to explicitly define the number of loops and runs, use the `-n` and `-r` options. Use `-n` to specify the loops and `-r` for the number the runs.



# Waterfall Charts: A Better Alternative to Line/Bar Plot



If you want to visualize a value over some period, a line (or bar) plot may not always be an apt choice.

A line-plot (or bar-plot) depicts the actual values in the chart. Thus, sometimes, it can get difficult to visually estimate the scale of incremental changes.

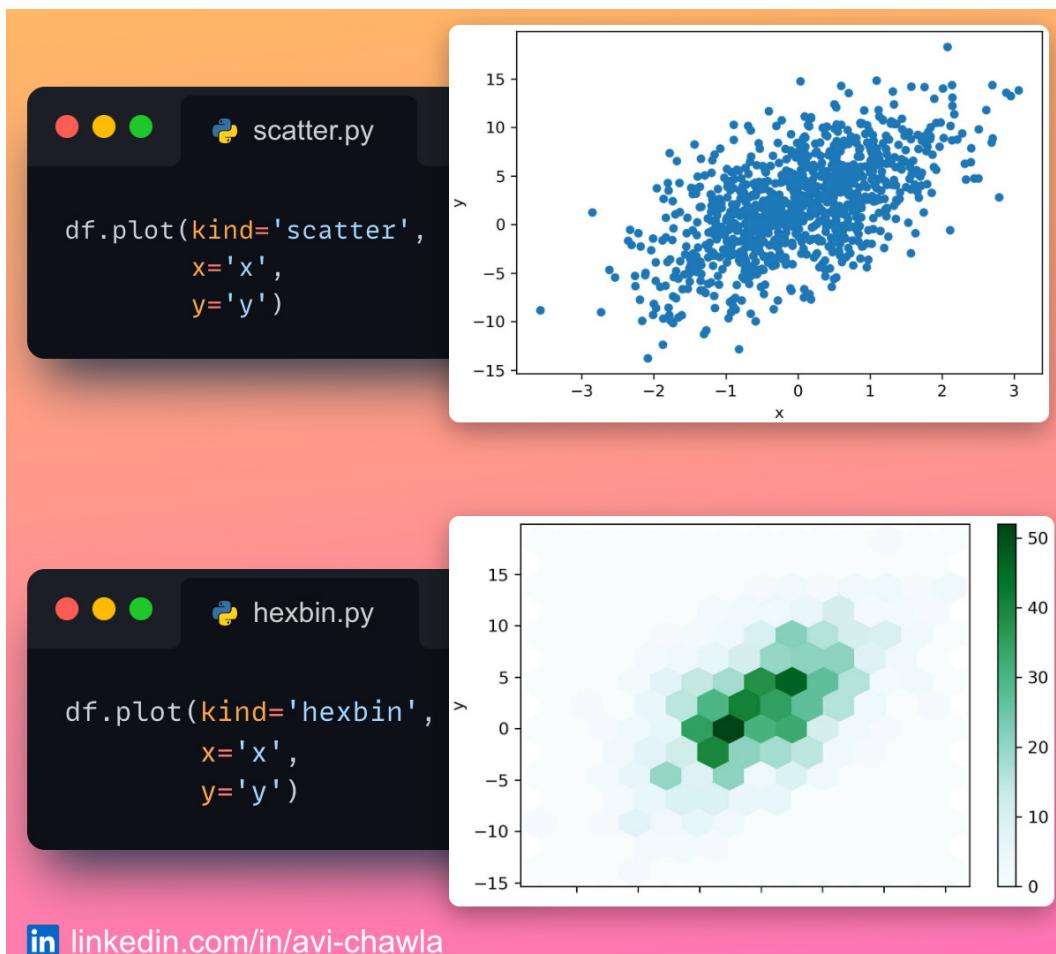
Instead, you can use a waterfall chart, which elegantly depicts these rolling differences.

To create one, you can use **waterfall\_chart** in Python. Here, the start and final values are represented by the first and last bars. Also, the marginal changes are automatically color-coded, making them easier to interpret.

Read more here: [GitHub](#).



# Hexbin Plots As A Richer Alternative to Scatter Plots



Scatter plots are extremely useful for visualizing two sets of numerical variables. But when you have, say, thousands of data points, scatter plots can get too dense to interpret.

Hexbins can be a good choice in such cases. As the name suggests, they bin the area of a chart into hexagonal regions. Each region is assigned a color intensity based on the method of aggregation used (the number of points, for instance).

Hexbins are especially useful for understanding the spread of data. It is often considered an elegant alternative to a scatter plot. Moreover, binning makes it easier to identify data clusters and depict patterns.



# Importing Modules Made Easy with Pyforest

The image shows two Jupyter notebook cells side-by-side. The top cell contains standard imports:import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
import sys  
  
from sklearn.linear\_model  
import LinearRegressionA white curved arrow points from the bottom cell to the first line of the top cell. The bottom cell shows the Pyforest import statement followed by four examples of modules imported with checkmarks:from pyforest import \*  
  
pd.read\_csv("file.csv") ✓  
np.array([1,2,3]) ✓  
sys.path ✓  
LinearRegression() ✓

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

The typical programming-related stuff in data science begins by importing relevant modules.

However, across notebooks/projects, the modules one imports are mostly the same. Thus, the task of importing all the individual libraries is kinda repetitive.

With **pyforest**, you can use the common Python libraries without explicitly importing them. A good thing is that it imports all the libraries with their standard conventions. For instance, **pandas** is imported with the **pd** alias.



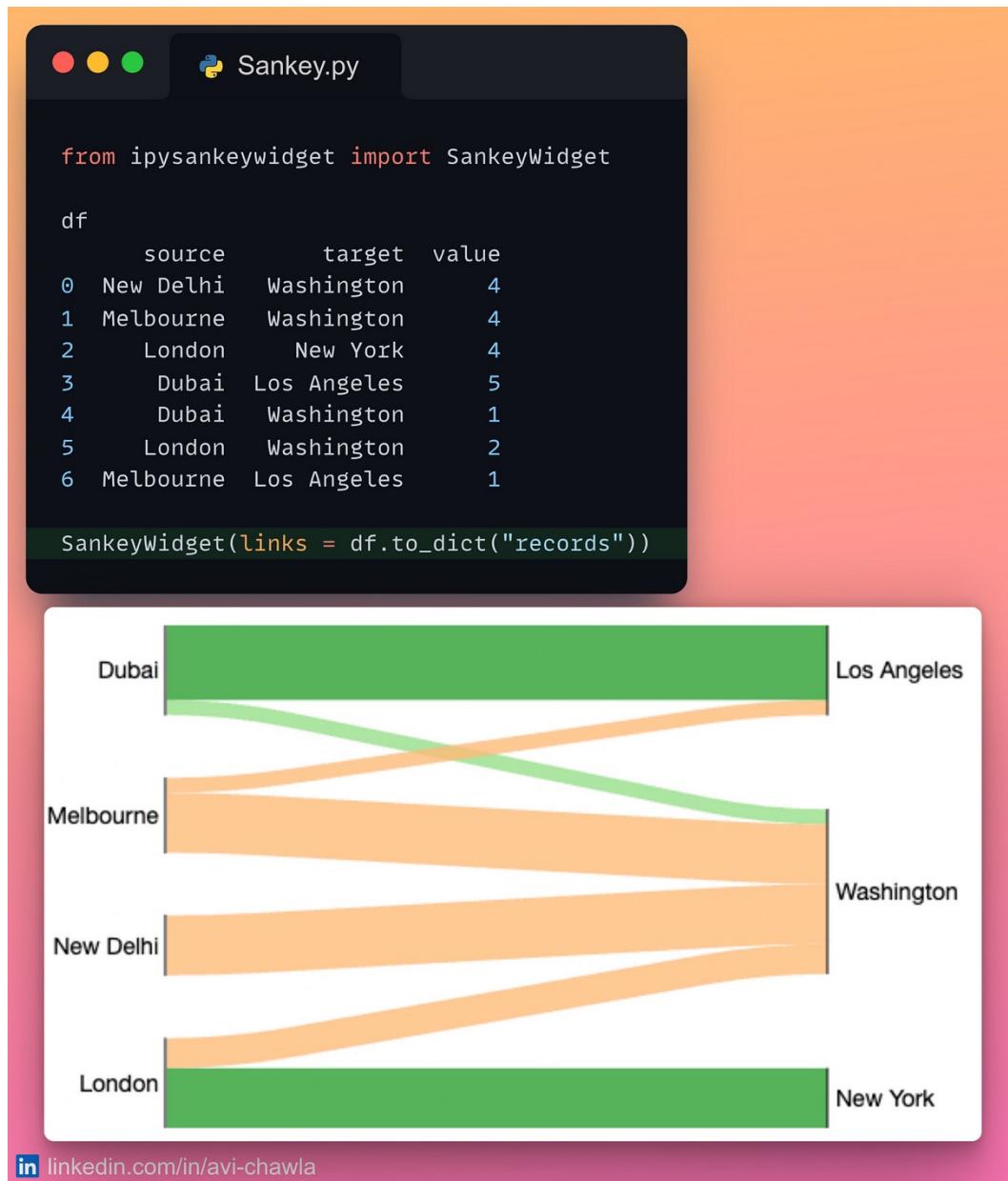
With that, you should also note that it is a good practice to keep Pyforest limited to prototyping stages. This is because once you say, develop and open-source your pipeline, other users may face some difficulties understanding it.

But if you are up for some casual experimentation, why not use it instead of manually writing all the imports?

Read more: [GitHub](#).



# Analyse Flow Data With Sankey Diagrams



Many tabular data analysis tasks can be interpreted as a flow between the source and a target.

Here, manually analyzing tabular reports/data to draw insights is typically not the right approach.

Instead, Flow diagrams serve as a great alternative in such cases.



Being visually appealing, they immensely assist you in drawing crucial insights from your data, which you may find challenging to infer by looking at the data manually.

For instance, from the diagram above, one can quickly infer that:

1. Washington hosts flights from all origins.
2. New York only receives passengers from London.
3. Majority of flights in Los Angeles come from Dubai.
4. All flights from New Delhi go to Washington.

Now imagine doing that by just looking at the tabular data. Not only will it be time-consuming, but there are chances that you may miss out on a few insights.

To generate a flow diagram, you can use floWeaver. It helps you to visualize flow data using Sankey diagrams.

Read more here: [Documentation](#).



# Feature Tracking Made Simple In Sklearn Transformers

The image shows two terminal windows on a Mac desktop. The left terminal window has a yellow background and displays the code for 'numpy\_output.py'. It includes a snippet of code that creates a DataFrame 'df' with columns 'col\_A' and 'col\_B', and then applies 'PolynomialFeatures().fit\_transform(df)' to it. The resulting output is a NumPy array:

	array([[ 1., 1., 2., 1., 2., 4.,
	[ 1., 3., 4., 9., 12., 16.],
	[ 1., 5., 6., 25., 30., 36.]])

A white callout box highlights this array output. A curved arrow points from this box up towards the right terminal window.

The right terminal window has a pink background and displays the code for 'pandas\_output.py'. It includes a snippet of code that uses 'set\_config(transform\_output = "pandas")' to configure the transformation, followed by the same DataFrame transformation code. The resulting output is a Pandas DataFrame:

	1	col_A	col_B	col_A^2	col_Acol_B	col_B^2
0	1.0	1.0	2.0	1.0	2.0	4.0
1	1.0	3.0	4.0	9.0	12.0	16.0
2	1.0	5.0	6.0	25.0	30.0	36.0

A white callout box highlights this DataFrame output. A curved arrow points from this box up towards the right terminal window.

At the bottom of the image, there is a blue LinkedIn icon followed by the URL 'linkedin.com/in/avi-chawla'.

Recently, [scikit-learn](#) announced the release of one of the most awaited improvements. In a gist, sklearn can now be configured to output Pandas DataFrames.

Until now, Sklearn's transformers were configured to accept a Pandas DataFrame as input. But they always returned a NumPy array as an output. As a result, the output had to be manually projected back to a



Pandas DataFrame. This, at times, made it difficult to track and assign names to the features.

For instance, consider the snippet above.

In **numpy\_output.py**, it is tricky to infer the name (or computation) of a column by looking at the NumPy array.

However, in the upcoming release, the transformer can return a Pandas DataFrame (**pandas\_output.py**). This makes tracking feature names incredibly simple.

Read more: [Release page](#).



# Lesser-known Feature of f-strings in Python

```
Count = 2
Fruit = "Apple"

print(f"Count = {Count}")
print(f"Fruit = {Fruit}")
## Count = 2
## Fruit = Apple
```

**Don't write  
variable name  
explicitly**

**Add "=" in  
curly braces {}** →

```
print(f"{Count = }")
print(f"{Fruit = }")
## Count = 2
## Fruit = Apple
```

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

While debugging, one often explicitly prints the name of the variable with its value to enhance code inspection.

Although there's nothing wrong with this approach, it makes your print statements messy and lengthy.

f-strings in Python offer an elegant solution for this.

To print the name of the variable, you can add an equals sign (=) in the curly braces after the variable. This will print the name of the variable along with its value but it is concise and clean.



# Don't Use `time.time()` To Measure Execution Time

The image shows two terminal windows on a Mac OS X desktop. The background has a blurred orange-to-red gradient.

The top window is titled "time.py" and contains the following Python code:

```
import time

start = time.time()
time.sleep(10)
end = time.time()

print(end - start)
## 10.00482
```

The bottom window is titled "perf\_counter.py" and contains the following Python code:

```
import time

start = time.perf_counter()
time.sleep(10)
end = time.perf_counter()

print(end - start)
## 10.00435
```

Below the windows, there is a blue LinkedIn icon followed by the URL "linkedin.com/in/avi-chawla".

The `time()` method from the `time` library is frequently used to measure the execution time.

However, `time()` is not meant for timing your code. Rather, its actual purpose is to tell the current time. This, at many times, compromises the accuracy of measuring the exact run time.

The correct approach is to use `perf_counter()`, which deals with relative time. Thus, it is considered the most accurate way to time your code.



# Now You Can Use DALL·E With OpenAI API

```
import openai

openai.api_key = "Your-API-Key"

response = openai.Image.create(
    prompt="The city of Paris on Mars.",
    n = 2
)

image_url = response['data'][0]['url']
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

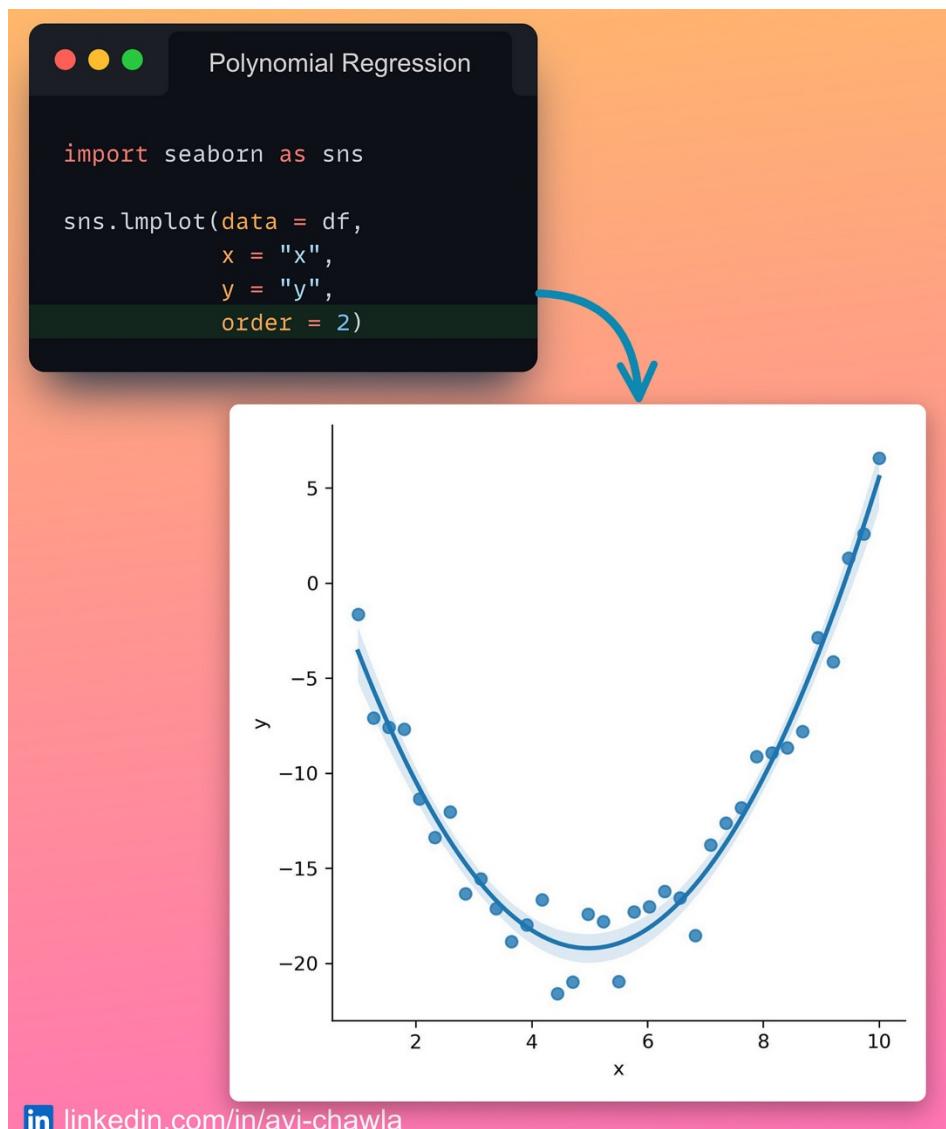
DALL·E is now accessible using the OpenAI API.

[\*\*OpenAI\*\*](#) recently made a big announcement. In a gist, developers can now integrate OpenAI's popular text-to-image model DALL·E into their apps using OpenAI API.

To achieve this, first, specify your API key (obtained after signup). Next, pass a text prompt to generate the corresponding image.



# Polynomial Linear Regression Plot Made Easy With Seaborn



[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

While creating scatter plots, one is often interested in displaying a linear regression (simple or polynomial) fit on the data points.

Here, training a model and manually embedding it in the plot can be a tedious job to do.

Instead, with Seaborn's **lmplot()**, you can add a regression fit to a plot, without explicitly training a model.

Specify the degree of the polynomial as the "**order**" parameter. Seaborn will add the corresponding regression fit on the scatter plot.

Read more here: [Seaborn Docs](#).



# Retrieve Previously Computed Output In Jupyter Notebook

```
In [3]: df.groupby("col1").col2.mean().reset_index()
Out[3]:
   col1  col2
0     A    4.0
1     B    3.0
2     C    5.0

In [4]: Out[3]
Out[4]:
   col1  col2
0     A    4.0
1     B    3.0
2     C    5.0
```

linkedin.com/in/avi-chawla

This is indeed one of the coolest things I have learned about Jupyter Notebooks recently.

Have you ever been in a situation where you forgot to assign the results obtained after some computation to a variable? Left with no choice, one has to unwillingly recompute the result and assign it to a variable for further use.

Thankfully, you don't have to do that anymore!

IPython provides a dictionary "**Out**", which you can use to retrieve a cell's output. All you need to do is specify the cell number as the dictionary's key, which will return the corresponding output. Isn't that cool?

View a video version of this post on LinkedIn: [Post Link](#).



# Parallelize Pandas Apply() With Swifter

```
Pandas Apply
df = ... ## Shape: (10M, 4)

def sum_row(row):
    return sum(row)

df.apply(sum_row, axis = 1)
```

Run-time:  
35 seconds

```
Swifter Apply
import swifter

df.swifter.apply(sum_row,
                axis = 1)
```

Run-time:  
15 seconds

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

The Pandas library has no inherent support to parallelize its operations. Thus, it always adheres to a single-core computation, even when other cores are idle.

Things get even worse when we use **apply()**. In Pandas, **apply()** is nothing but a glorified for-loop. As a result, it cannot even take advantage of vectorization.

A quick solution to parallelize **apply()** is to use **swifter** instead.

Swifter allows you to apply any function to a Pandas DataFrame in a parallelized manner. As a result, it provides considerable performance gains while preserving the old syntax. All you have to do is use **df.swifter.apply** instead of **df.apply**.

Read more here: [Swifter Docs](#).



# Create DataFrame Hassle-free By Using Clipboard

The screenshot illustrates a two-step process for creating a DataFrame from tabular data:

**Step 1: Copy Table** (Left): A Jupyter Notebook cell titled "Products" contains Python code to filter a DataFrame for rows where column A equals "foo". It prints the resulting DataFrame, which has columns A, B, C, and D, and rows 0 through 7. The output shows values like "foo" in column A and 0 in column C for row 0.

**Step 2: Read in Pandas** (Right): A terminal window titled "Read Clipboard" shows the command `df = pd.read_clipboard()`. Below it, the command `>>> df.head()` is run, displaying the same DataFrame as in Step 1.

LinkedIn link: [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Many Pandas users think that a DataFrame can ONLY be loaded from disk. However, this is not true.

Imagine one wants to create a DataFrame from tabular data printed on a website. Here, they are most likely to be tempted to copy the contents to a CSV and read it using Pandas' `read_csv()` method. But this is not an ideal approach here.

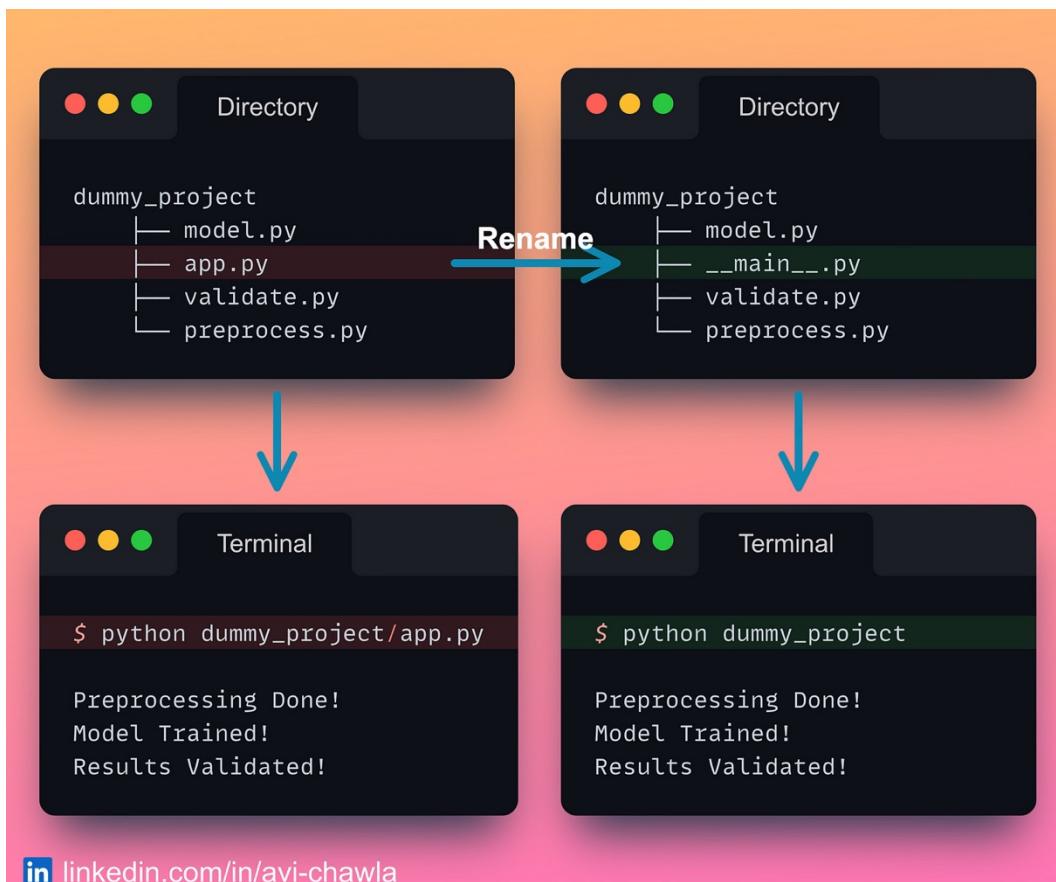
Instead, with the `read_clipboard()` method, you can eliminate the CSV step altogether.

This method allows you to create a DataFrame from tabular data stored in a clipboard buffer. Thus, you just need to copy the data and invoke the method to create a DataFrame. This is an elegant approach that saves plenty of time.

Read more here: [Pandas Docs](#).



# Run Python Project Directory As A Script



A Python script is executed when we run a **.py** file. In large projects with many files, there's often a source (or base) Python file we begin our program from.

To make things simpler, you can instead rename this base file to **\_\_main\_\_.py**. As a result, you can execute the whole pipeline by running the parent directory itself.

This is concise and also makes it slightly easier for other users to use your project.



# Inspect Program Flow with IceCream

The diagram illustrates the use of the IceCream library for inspecting program flow. It consists of two main sections, each with a code editor window and a terminal window.

**Top Example:**

```
file.py
1 def func():
2     print(0)
3     ...
4
5     if condition:
6         print(1)
7         ...
8     else:
9         print(2)
10    ...
```

**Terminal Output:**

```
$ python file.py
0
2
```

**Bottom Example:**

```
file.py
1 from icecream import ic
2
3 def func():
4     ic()
5     ...
6     if condition:
7         ic()
8         ...
9     else:
10        ic()
11    ...
```

**Terminal Output:**

```
$ python file.py
ic| file.py:4 in func()
ic| file.py:10 in func()
```

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

While debugging, one often writes many `print()` statements to inspect the program's flow. This is especially true when we have many IF conditions.

Using empty `ic()` statements from the IceCream library can be a better alternative here. It outputs many additional details that help in inspecting the flow of the program.

This includes the line number, the name of the function, the file name, etc.

Read more in my Medium Blog: [Link](#).



# Don't Create Conditional Columns in Pandas with Apply

```
def assign_class(num):  
  
    if num>0.5:  
        return "Class A"  
    else:  
        return "Class B"  
  
df.col1.apply(assign_class)  
## 987 ms ± 47.1 ms per loop
```

```
import numpy as np  
  
np.where(df["col1"]>0.5,  
         "Class A",  
         "Class B")  
## 194 ms ± 23.7 ms per loop
```

If condition

is True

If condition

is False

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

While creating conditional columns in Pandas, we tend to use the `apply()` method almost all the time.

However, `apply()` in Pandas is nothing but a glorified for-loop. As a result, it misses the whole point of vectorization.

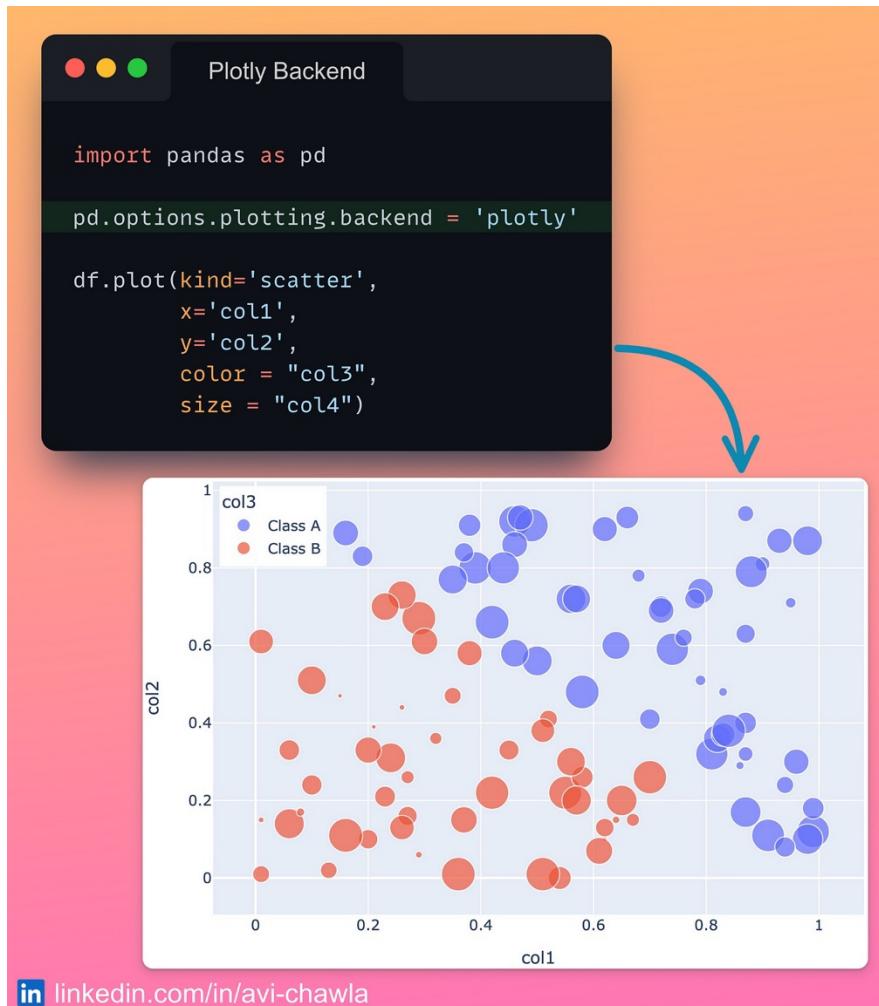
Instead, you should use the `np.where()` method to create conditional columns. It does the same job but is extremely fast.

The condition is passed as the first argument. This is followed by the result if the condition evaluates to True (second argument) and False (third argument).

Read more here: [NumPy docs](#).



# Pretty Plotting With Pandas



Matplotlib is the default plotting API of Pandas. This means you can create a Matplotlib plot in Pandas, without even importing it.

Despite that, these plots have always been basic and not so visually appealing. Plotly, with its pretty and interactive plots, is often considered a suitable alternative. But familiarising yourself with a whole new library and its syntax can be time-consuming.

Thankfully, Pandas does allow you to change the default plotting backend. Thus, you can leverage third-party visualization libraries for plotting with Pandas. This makes it effortless to create prettier plots while almost preserving the old syntax.



# Build Baseline Models Effortlessly With Sklearn

```
from sklearn.dummy import DummyClassifier

dummy_clf = DummyClassifier(
    strategy="most_frequent"
).fit(X, y)

>>> dummy_clf.predict(X)
array([0, 0, 0, 0, 0])

>>> dummy_clf.score(X, y)
0.6
```

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

Before developing a complex ML model, it is always sensible to create a baseline first.

The baseline serves as a benchmark for the engineered model. Moreover, it ensures that the model is better than making random (or fixed) predictions. But building baselines with various strategies (random, fixed, most frequent, etc.) can be tedious.

Instead, Sklearn's **DummyClassifier()** (and **DummyRegressor()**) makes it totally effortless and straightforward. You can select the specific behavior of the baseline with the **strategy** parameter.

Read more here: [Documentation](#).



# Fine-grained Error Tracking With Python 3.11

```
$ python expt.py

Traceback (most recent call last):

  File "expt.py", line 11, in <module>
    print(function(a=2, b=0))
                       ^^^^^^^^^^

  File "expt.py", line 6, in function
    return (b / a) + (a / b)
                      ~~^~~

ZeroDivisionError: division by zero
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Python 3.11 was released today, and many exciting features have been introduced.

For instance, various speed improvements have been implemented. As per the official release, Python 3.11 is, on average, 25% faster than Python 3.10. Depending on your work, it can be up to 10-60% faster.

One of the coolest features is the fine-grained error locations in tracebacks.

In Python 3.10 and before, the interpreter showed the specific line that caused the error. This, at many times, caused ambiguity during debugging.

In Python 3.11, the interpreter will point to the exact location that caused the error. This will immensely help programmers during debugging.

Read more here: [Official Release](#).



# Find Your Code Hiding In Some Jupyter Notebook With Ease



Programmers who use Jupyter often refer to their old notebooks to find a piece of code.

However, it gets tedious when they have multiple files to look for and can't recall the specific notebook of interest. The file name

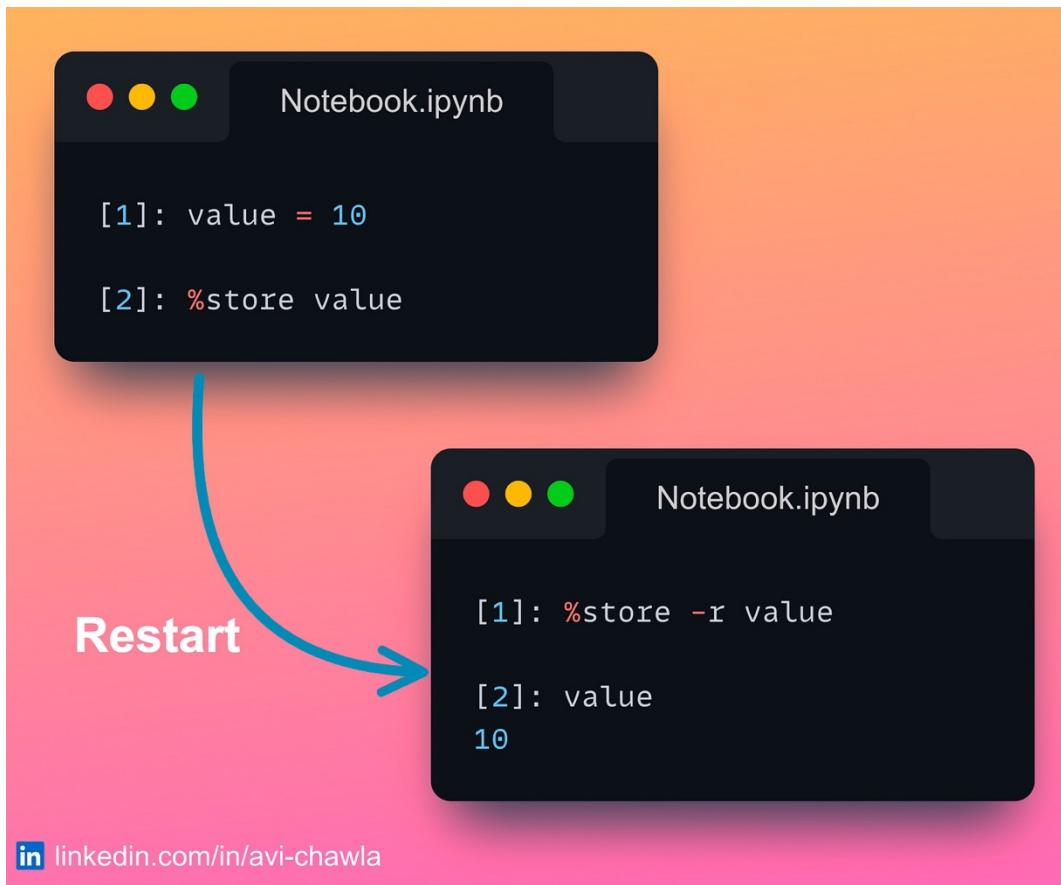
**Untitled1.ipynb**, ..., and **Untitled82.ipynb**, don't make it any easier.

The "**grep**" command is a much better solution to this. Very know that you can use "**grep**" in the command line to search in notebooks, as you do in other files (.txt, for instance). This saves plenty of manual work and time.

P.S. How do you find some previously written code in your notebooks (if not manually)?



# Restart the Kernel Without Losing Variables



While working in a Jupyter Notebook, you may want to restart the kernel due to several reasons. But before restarting, one often tends to dump data objects to disk to avoid recomputing them in the subsequent run.

The "store" magic command serves as an ideal solution to this. Here, you can obtain a previously computed value even after restarting your kernel. What's more, you never need to go through the hassle of dumping the object to disk.



# How to Read Multiple CSV Files Efficiently

```
import pandas as pd

files = ["jan.csv", "feb.csv",
         "mar.csv", "apr.csv",
         "may.csv", "jun.csv"]
## 300 MBs each

df_list = []
for i in files:
    df_list.append(pd.read_csv(i))
data = pd.concat(df_list)
```

Run-time: 64 seconds

```
import datatable as dt

files = ["jan.csv", "feb.csv",
         "mar.csv", "apr.csv",
         "may.csv", "jun.csv"]
## 300 MBs each

df = dt.iread(files) ## read files
df = dt.rbind(df) ## concatenate row-wise
df = df.to_pandas() ## convert to Pandas
```

Run-time: 36 seconds

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

In many situations, the data is often split into multiple CSV files and transferred to the DS/ML team for use.

As Pandas does not support parallelization, one has to iterate over the list of files and read them one by one for further processing.

"Datatable" can provide a quick fix for this. Instead of reading them iteratively with Pandas, you can use Datatable to read a bunch of files. Being parallelized, it provides a significant performance boost as compared to Pandas.



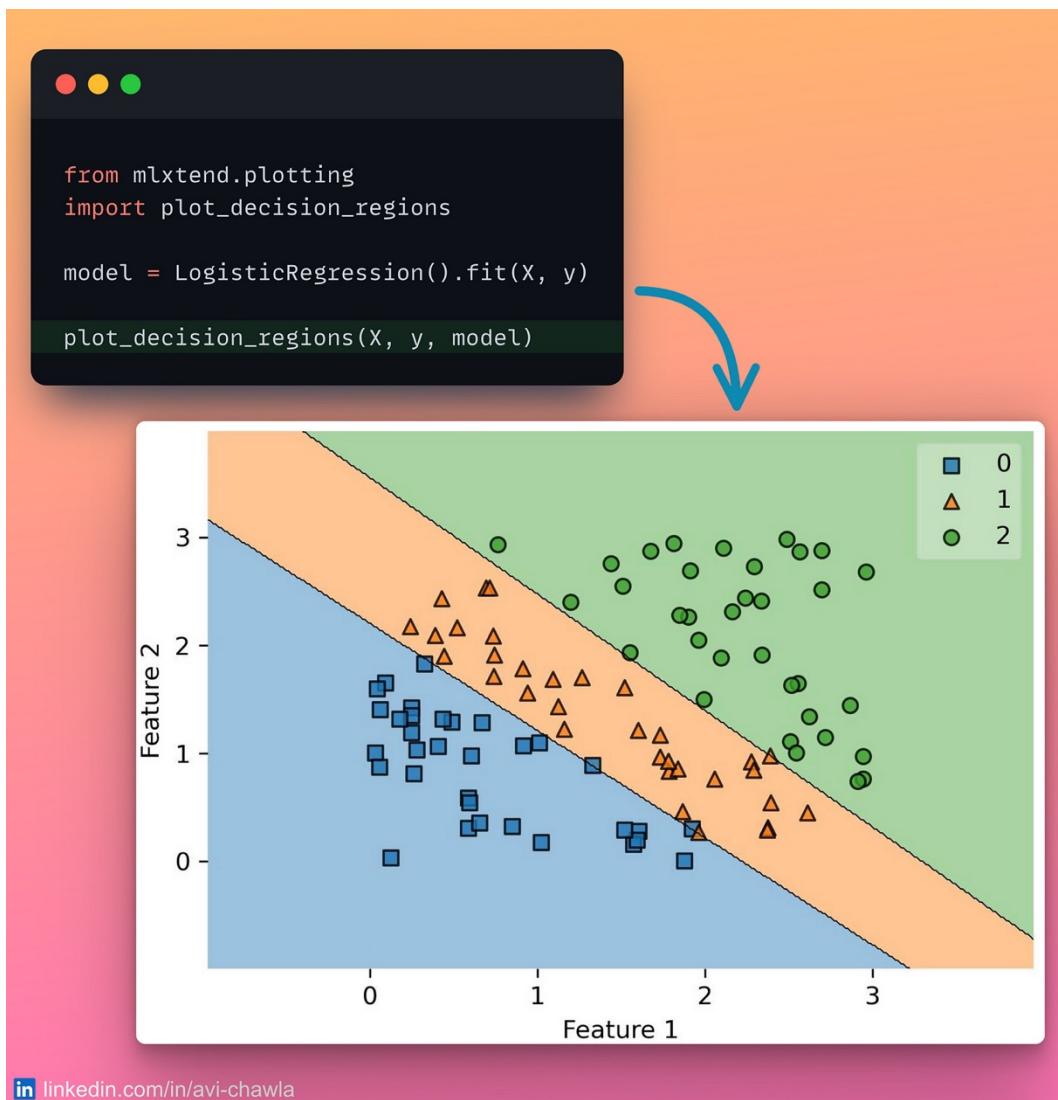
[avichawla.substack.com](https://avichawla.substack.com)

The performance gain is not just limited to I/O but is observed in many other tabular operations as well.

Read more here: [DataTable Docs.](#)



# Elegantly Plot the Decision Boundary of a Classifier



[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Plotting the decision boundary of a classifier can reveal many crucial insights about its performance.

Here, region-shaded plots are often considered a suitable choice for visualization purposes. But, explicitly creating one can be extremely time-consuming and complicated.

MLxtend condenses that to a simple one-liner in Python. Here, you can plot the decision boundary of a classifier with ease, by just providing it the model and the data.



# An Elegant Way to Import Metrics From Sklearn

The image shows two Jupyter notebook environments side-by-side. The left notebook illustrates the traditional method of importing multiple metrics from `sklearn.metrics`, while the right notebook demonstrates an elegant alternative using `get_scoring`.

**Left Notebook (Traditional Import):**

```
from sklearn.metrics
import accuracy_score, f1_score,
precision_score, recall_score,
roc_auc_score, ...

>>> accuracy_score(y_true, y_pred)
0.5

>>> precision_score(y_true, y_pred)
0.8
```

**Right Notebook (Elegant Import):**

```
from sklearn.metrics import get_scoring

accuracy = get_scoring("accuracy")
>>> accuracy._score_func(y_true, y_pred)
0.5

precision = get_scoring("precision")
>>> precision._score_func(y_true, y_pred)
0.8
```

A blue arrow points from the text "Import all metrics individually" to the first code block. A blue curved arrow points from the text "Get a scorer from string" to the second code block.

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

While using **scikit-learn**, one often imports multiple metrics to evaluate a model. Although there is nothing wrong with this practice, it makes the code inelegant and cluttered - with the initial few lines of the file overloaded with imports.

Instead of importing the metrics individually, you can use the `get_scoring()` method. Here, you can pass the metric's name as a string, and it returns a scorer object for you.

Read more here: [Scikit-learn page](#).



# Configure Sklearn To Output Pandas DataFrame

The image shows two Jupyter notebook cells side-by-side. The top cell, titled 'Scikit-learn 1.1', contains Python code for scaling a Pandas DataFrame:from sklearn.preprocessing  
import StandardScaler  
  
X\_train = ... ## Pandas DataFrame  
  
scaler = StandardScaler()  
X\_scaled = scaler.fit\_transform(X\_train)  
  
type(X\_scaled) ## numpy.ndarrayA callout bubble to the right of this cell says 'Output is NumPy Array'. The bottom cell, titled 'Scikit-learn 1.2.dev', shows the same code but includes a call to 'set\_output' to specify Pandas output:scaler = StandardScaler()  
scaler.set\_output(transform="pandas")  
X\_scaled = scaler.fit\_transform(X\_train)  
  
type(X\_scaled) ## pandas.core.frame.DataFrameA callout bubble to the left of this cell says 'Output is Pandas DataFrame'. A curved arrow points from the 'NumPy Array' callout down to the 'Pandas DataFrame' callout.

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

Recently, Scikit-learn announced the release of one of the most awaited improvements. In a gist, sklearn can now be configured to output Pandas DataFrames instead of NumPy arrays.

Until now, Sklearn's transformers were configured to accept a Pandas DataFrame as input. But they always returned a NumPy array as an output. As a result, the output had to be manually projected back to a Pandas DataFrame.

Now, the **set\_output** API will let transformers output a Pandas DataFrame instead.

This will make running pipelines on DataFrames smoother. Moreover, it will provide better ways to track feature names.



# Display Progress Bar With Apply() in Pandas

The image shows a Jupyter Notebook interface with two code cells side-by-side. A blue arrow points from the first cell to the second.

**Without Progress:**

```
import pandas as pd  
df.apply(func)
```

**With Progress:**

```
import pandas as pd  
from tqdm.notebook import tqdm  
tqdm.pandas()  
a = df.progress_apply(func)
```

A progress bar at the bottom indicates 100% completion with the text: 1000000/1000000 [00:04<00:00, 281929.55it/s].

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

While applying a method to a DataFrame using **apply()**, we don't get to see the progress and an estimated remaining time.

To resolve this, you can instead use **progress\_apply()** from **tqdm** to display a progress bar while applying a method.

Read more here: [GitHub](#).



# Modify a Function During Run-time

The screenshot shows two code snippets in a Jupyter Notebook cell. The first snippet contains a function `func` that prints odd numbers from 1 to 19. The second snippet shows the same function with an additional print statement for even numbers. A red arrow points from the first snippet to the second, labeled "Modified During Run-time". Below the code, the output of the cell is shown in two parts, each enclosed in a box. The left box shows the output for odd numbers (1 to 19), and the right box shows the output for even numbers (2 to 18). A LinkedIn link is at the bottom.

```
from time import sleep
from reloading import reloading

@reloading
def func(num):
    if number % 2:
        print(f"{number} is Odd")
    else:
        pass

for i in range(100):
    func(i)
```

Modified During Run-time →

```
from time import sleep
from reloading import reloading

@reloading
def func(num):
    if number % 2:
        print(f"{number} is Odd")
    else:
        print(f"{number} is Even")

for i in range(100):
    func(i)
```

```
1 is Odd
3 is Odd
5 is Odd
7 is Odd
9 is Odd
11 is Odd
13 is Odd
14 is Even
15 is Odd
16 is Even
17 is Odd
18 is Even
19 is Odd
```

```
2 is Even
4 is Even
6 is Even
8 is Even
10 is Even
12 is Even
14 is Even
16 is Even
18 is Even
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Have you ever been in a situation where you wished to add more details to an already running code?

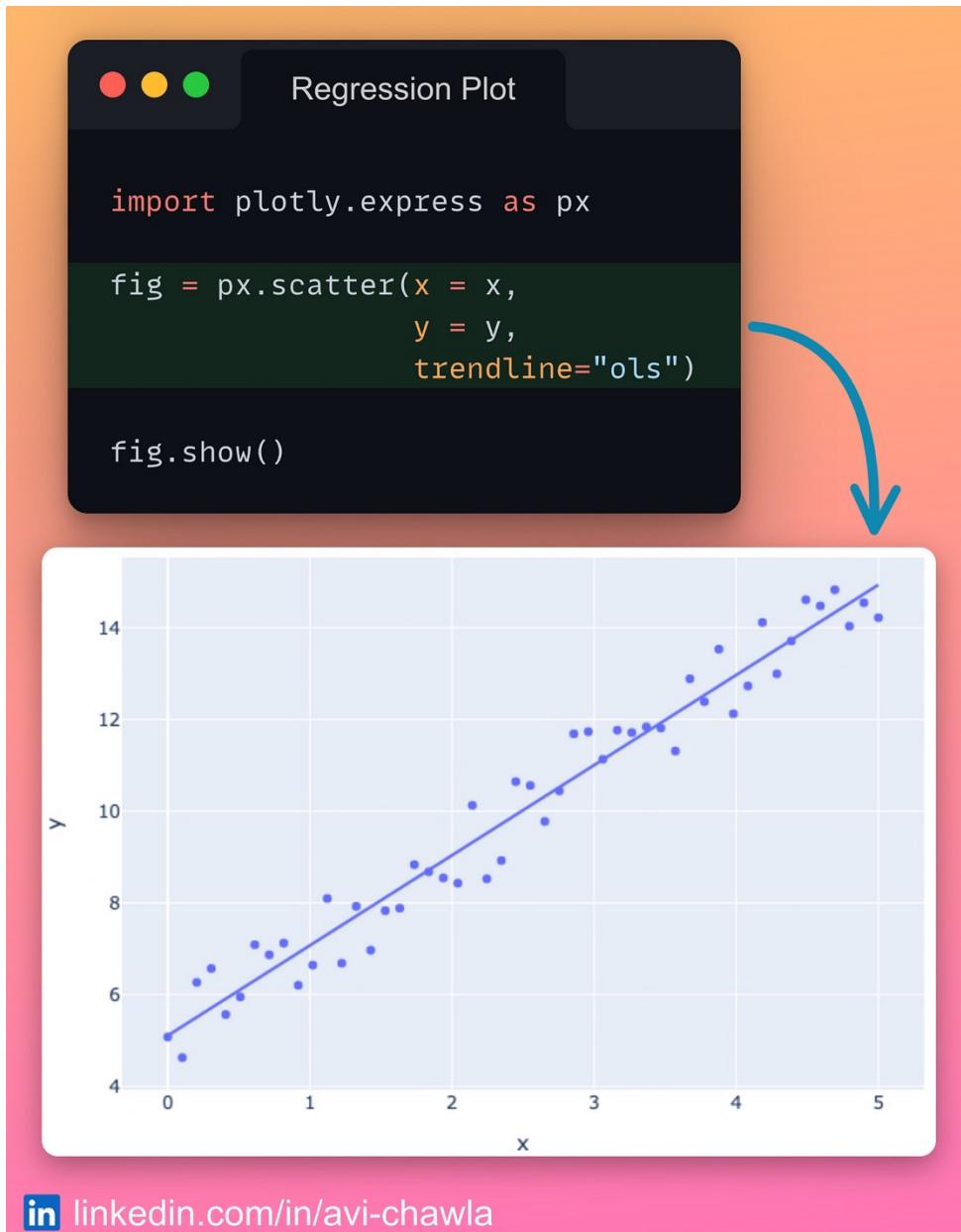
This is typically observed in ML where one often forgets to print all the essential training details/metrics. Executing the entire code again, especially when it has been up for some time is not an ideal approach here.

If you want to modify a function during execution, decorate it with the `reloading` decorator (**@reloading**). As a result, Python will reload the function from the source before each execution.

Link to reloading: [GitHub](#).



# Regression Plot Made Easy with Plotly



[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

While creating scatter plots, one is often interested in displaying a simple linear regression fit on the data points.

Here, training a model and manually embedding it in the plot can be a tedious job to do.

Instead, with Plotly, you can add a regression line to a plot, without explicitly training a model.

Read more [here](#).



# Polynomial Linear Regression with NumPy

The image shows two Jupyter notebook cells side-by-side. The top cell is titled 'Sklearn' and the bottom cell is titled 'NumPy'. Both cells contain Python code for polynomial regression.

**Sklearn Cell:**

```
## 1 Degree Polynomial  
model = LinearRegression().fit(x, y)  
  
## 2 Degree Polynomial  
x = np.hstack((x, x**2))  
model = LinearRegression().fit(x, y)  
  
>>> x = 2  
>>> inp = np.array([[x, x**2]])  
>>> model.predict(inp)  
-10.4
```

**NumPy Cell:**

```
coeff = np.polyfit(x, y, deg = 2)  
model = np.poly1d(coeff)  
  
>>> inp = 2  
>>> model(inp)  
-10.4
```

Annotations with arrows point from the explanatory text below to specific parts of the code:

- An arrow points to the line `x = np.hstack((x, x**2))` in the Sklearn cell with the text "Create Polynomial Features".
- An arrow points to the line `deg = 2` in the NumPy cell with the text "Specify Degree".

Polynomial linear regression using Sklearn is tedious as one has to explicitly code its features. This can get challenging when one has to iteratively build higher-degree polynomial models.

NumPy's **polyfit()** method is an excellent alternative to this. Here, you can specify the degree of the polynomial as a parameter. As a result, it automatically creates the corresponding polynomial features.

The downside is that you cannot add custom features such as trigonometric/logarithmic. In other words, you are restricted to only polynomial features. But if that is not your requirement, NumPy's **polyfit()** method can be a better approach.

Read more:

<https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html>.



# Alter the Datatype of Multiple Columns at Once

The screenshot shows two Jupyter Notebook cells. The top cell displays a DataFrame with four columns: col1, col2, col3, and col4. The bottom cell shows three separate `astype` calls to convert the datatypes of these columns.

```
>>> df
   col1  col2  col3  col4
0     1     7     4     A
1     3     9     6     B
2     6     2     5     A
```

```
df["col1"] = df.col1.astype(np.int32)
df["col2"] = df.col2.astype(np.int16)
df["col3"] = df.col3.astype(np.float16)
```

A blue arrow points from the text "Multiple Calls" to the bottom cell. Another blue arrow points from the text "Single Call" to the bottom cell, indicating a comparison between the two methods.

Single Call →

```
df = df.astype({
    "col1":np.int32,
    "col2":np.int16,
    "col3":np.float16})
```

in [linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

A common approach to alter the datatype of multiple columns is to invoke the `astype()` method individually for each column.

Although the approach works as expected, it requires multiple function calls and more code. This can be particularly challenging when you want to modify the datatype of many columns.

As a better approach, you can condense all the conversions into a single function call. This is achieved by passing a dictionary of column-to-datatype mapping, as shown below.



# Datatype For Handling Missing Valued Columns in Pandas

The screenshot shows two code snippets in a Jupyter Notebook environment.

**NaN column:**

```
>>> len(df.col1)
## Total entries: 1,000,000

>>> len(df[df.col1.isna()])
## NaN entries: 700,000 (70%)
```

**Sparse Datatype:**

```
df.col1.memory_usage()
## Memory usage before conversion: 7.6 MB

df["col1"] = df.col1.astype("Sparse[float32]")

df.col1.memory_usage()
## Memory usage after conversion: 2.0 MB
```

[in linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

If your data has NaN-valued columns, Pandas provides a datatype specifically for representing them - called the Sparse datatype.

This is especially handy when you are working with large data-driven projects with many missing values.

The snippet compares the memory usage of float and sparse datatype in Pandas.



# Parallelize Pandas with Pandarallel

```
Python Pandarallel.py

from pandarallel import pandarallel
pandarallel.initialize()

def add_row(row):
    return sum(row)

df = ... ## 10M Rows, 2 Columns
```

```
Apply vs Parallel Apply

df.apply(add_row, axis = 1)
## 53 secs

df.parallel_apply(add_row, axis = 1)
## 11 secs
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Pandas' operations do not support parallelization. As a result, it adheres to a single-core computation, even when other cores are available. This makes it inefficient and challenging, especially on large datasets.

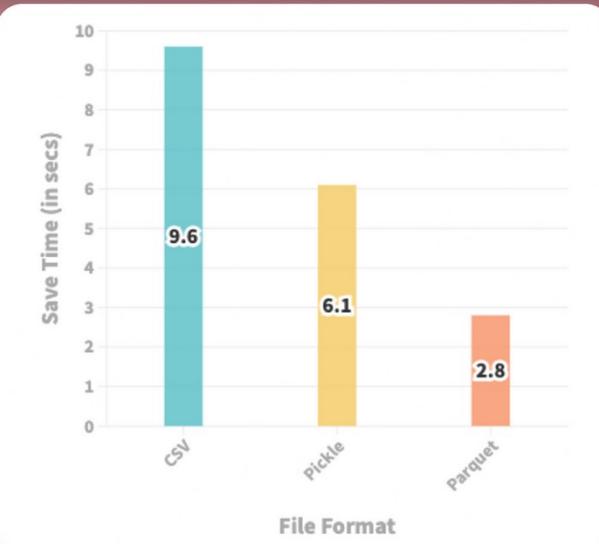
"Pandarallel" allows you to parallelize its operations to multiple CPU cores - by changing just one line of code. Supported methods include `apply()`, `applymap()`, `groupby()`, `map()` and `rolling()`.

Read more: [GitHub](#).



# Why you should not dump DataFrames to a CSV

```
1 df = ... ## 1M Rows, 30 Columns
2
3 df.to_csv("file.csv")
4
5 df.to_pickle("file.pickle")
6
7 df.to_parquet("file.parquet")
```



[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

The CSV file format is widely used to save Pandas DataFrames. But are you aware of its limitations? To name a few,

1. The CSV does not store the datatype information. Thus, if you modify the datatype of column(s), save it to a CSV, and load again, Pandas will not return the same datatypes.
2. Saving the DataFrame to a CSV file format isn't as optimized as



other supported formats by Pandas. These include Parquet, Pickle, etc.

Of course, if you need to view your data outside Python (Excel, for instance), you are bound to use a CSV. But if not, prefer other file formats.

Further reading: [Why I Stopped Dumping DataFrames to a CSV and Why You Should Too.](#)



# Save Memory with Python Generators

The image shows two side-by-side terminal windows. The left window is titled 'List.py' and the right is titled 'Generator.py'. Both windows run the same code, which uses the `sys.getsizeof` function to measure the memory footprint of a large iterable. In 'List.py', a list is created, resulting in a memory usage of 89095160 bytes. In 'Generator.py', a generator is created using a comprehension, resulting in a memory usage of only 112 bytes. This demonstrates that generators are more memory-efficient than lists for large static iterables.

```
1 from sys import getsizeof
2
3 my_list = [i for i in range(10**7)]
4 ## use [] to create a list
5
6 >>> getsizeof(my_list)
7 ## 89095160 bytes
8
9 >>> sum(my_list)
10 ## 49999995000000
11
12 >>> sum(my_list)
13 ## 49999995000000
```

```
1 from sys import getsizeof
2
3 my_gen = (i for i in range(10**7))
4 ## use () to create a generator
5
6 >>> getsizeof(my_gen)
7 ## 112 bytes
8
9 >>> sum(my_gen)
10 ## 49999995000000
11
12 >>> sum(my_gen)
13 ## 0
```

If you use large static iterables in Python, a list may not be an optimal choice, especially in memory-constrained applications.

A list stores the entire collection in memory. However, a generator computes and loads a single element at a time **ONLY** when it is required. This saves both memory and object creation time.

Of course, there are some limitations of generators too. For instance, you cannot use common list operations such as `append()`, `slicing`, etc.

Moreover, every time you want to reuse an element, it must be regenerated (see `Generator.py`: line 12).



## Don't use `print()` to debug your code.

```
Print

def func(arr, n):
    print("arr =", arr, "n =", n)

func([1,2,3], 2)
## arr = [1, 2, 3] n = 2
```

```
Icecream

from icecream import ic

def func(arr, n):
    ic(arr, n)

func([1,2,3], 2)
## ic| arr: [1, 2, 3], n: 2
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

Debugging with `print` statements is a messy and inelegant approach. It is confusing to map the output to its corresponding debug statement. Moreover, it requires extra manual formatting to comprehend the output.

The "**icecream**" library in Python is an excellent alternative to this. It makes debugging effortless and readable, with minimal code.



[avichawla.substack.com](https://avichawla.substack.com)

Features include printing expressions, variable names, function names, line numbers, filenames, and many more.

P.S. The snippet only gives a brief demonstration. However, the actual functionalities are much more powerful and elegant as compared to debugging with `print()`.

More about icecream

here: <https://github.com/gruns/icecream>.



# Find Unused Python Code With Ease

```
code.py
```

```
1 def sum_func(arr):
2     return sum(arr)
3
4 def max_func(arr):
5     return max(arr)
6
7 if __name__ == "__main__":
8
9     input_arr = [1, 3, 5, 2, 9]
10    flag = 1
11
12    input_sum = sum_func(input_arr)
13    print(input_sum)
```

```
Terminal
```

```
$ vulture code.py
```

```
code.py:4: unused function 'max_func'
code.py:10: unused variable 'flag'
```

[linkedin.com/in/avi-chawla](https://www.linkedin.com/in/avi-chawla)

As the size of your codebase increases, so can the number of instances of unused code. This inhibits its readability and conciseness.

With the "vulture" module in Python, you can locate dead (unused) code in your pipeline, as shown in the snippet.



# Define the Correct DataType for Categorical Columns

```
import pandas as pd

len(df.Gender)
## 1500

df.Gender.unique()
## ["Male", "Female"]
```

```
Reduce Memory Usage

import pandas as pd

df.Gender.memory_usage(), df.Gender.dtype
## 90.5 KB, object

df["Gender"] = df.Gender.astype("category")

df.Gender.memory_usage(), df.Gender.dtype
## 1.8 KB, CategoricalDtype
```

[in](https://www.linkedin.com/in/avi-chawla) linkedin.com/in/avi-chawla

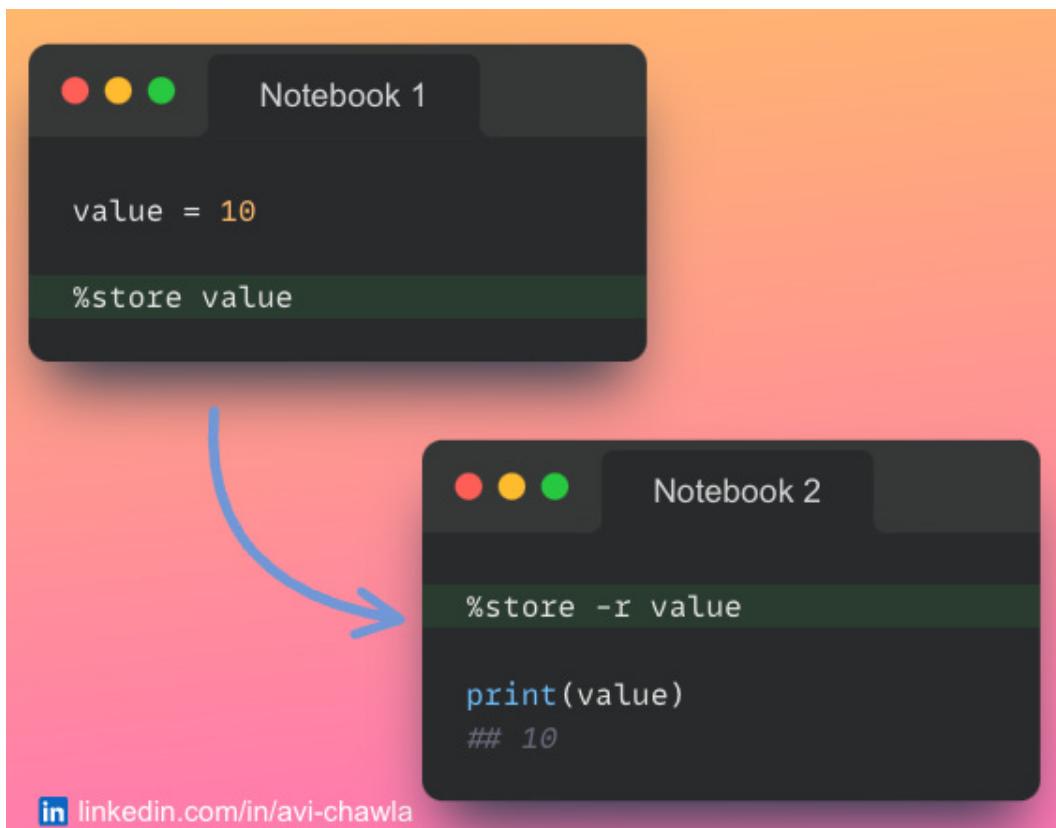
If your data has categorical columns, you should not represent them as int/string data type.

Rather, Pandas provides an optimized data type specifically for categorical columns. This is especially handy when you are working with large data-driven projects.

The snippet compares the memory usage of string and categorical data types in Pandas.



# Transfer Variables Between Jupyter Notebooks



While working with multiple jupyter notebooks, you may need to share objects between them.

With the "store" magic command, you can transfer variables across notebooks without storing them on disk.

P.S. You can also restart the kernel and retrieve an old variable with "store".



# Why You Should Not Read CSVs with Pandas

```
Pandas
1 file = "file.csv"
2 ## 1M rows and 30 columns
3
4 import pandas as pd
5
6 df = pd.read_csv(file)
7 ## 8.82 secs

Datatable
1 file = "file.csv"
2
3 import datatable as dt
4
5 df = dt.fread(file)
6 df = df.to_pandas()
7 ## 4.04 secs (line 5 + 6)
```

Pandas adheres to a single-core computation, which makes its operations extremely inefficient, especially on large datasets.

The "datatable" library in Python is an excellent alternative with a Pandas-like API. Its multi-threaded data processing support makes it faster than Pandas.

The snippet demonstrates the run-time comparison of creating a "Pandas DataFrame" from a CSV using Pandas and Datatable.



# Modify Python Code During Run-Time

The diagram illustrates the process of modifying Python code while it is running. It consists of two terminal windows and a status bar.

**Top Terminal Window:**

```
from time import sleep
from reloading import reloading

for number in reloading(range(100)):

    if number % 2:
        print(f"{number} is Odd")
    else:
        pass
```

**Output:**

```
1 is Odd
3 is Odd
5 is Odd
7 is Odd
9 is Odd
11 is Odd
```

**Status Bar:**

Modified During Run-time

**Bottom Terminal Window:**

```
from time import sleep
from reloading import reloading

for number in reloading(range(100)):

    if number % 2:
        print(f"{number} is Odd")
    else:
        print(f"{number} is Even")
```

**Output:**

```
1 is Odd
3 is Odd
5 is Odd
7 is Odd
9 is Odd
11 is Odd
13 is Odd
14 is Even
15 is Odd
16 is Even
17 is Odd
18 is Even
19 is Odd
```

**LinkedIn Link:**

[linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

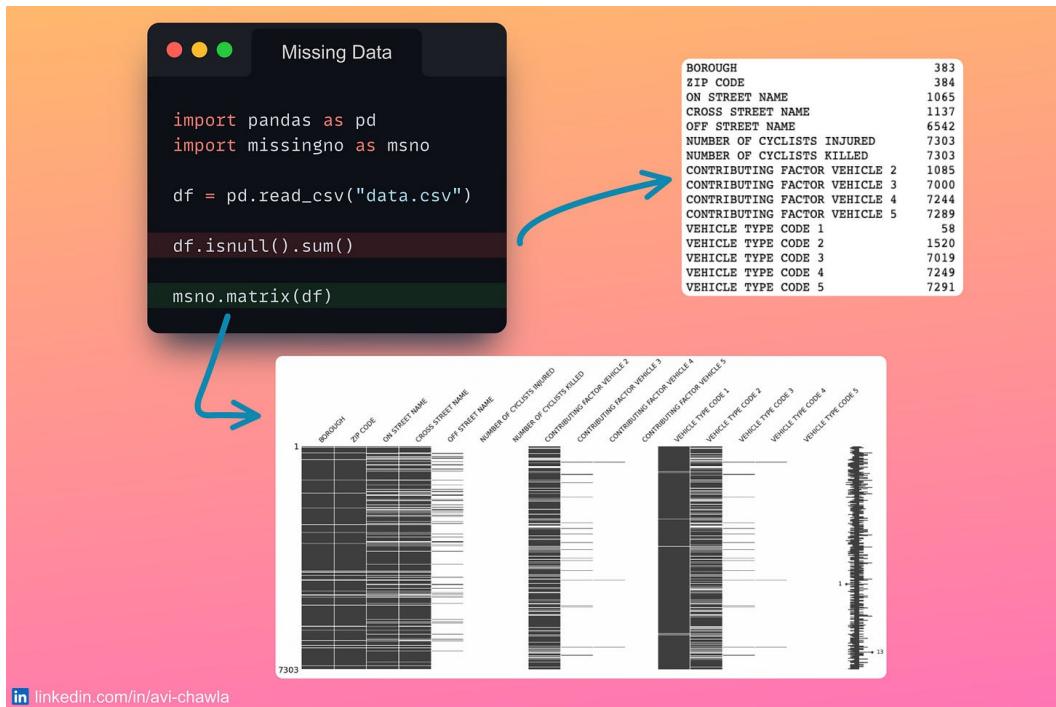
Have you ever been in a situation where you wished to add more details to an already running code (printing more details in a for-loop, for instance)?

Executing the entire code again, especially when it has been up for some time, is not the ideal approach here.

With the "reloading" library in Python, you can add more details to a running code without losing any existing progress.



# Handle Missing Data With Missingno



If you want to analyze missing values in your dataset, Pandas may not be an apt choice.

Pandas' methods hide many important details about missing values. These include their location, periodicity, the correlation across columns, etc.

The "missingno" library in Python is an excellent resource for exploring missing data. It generates informative visualizations for improved data analysis.

The snippet demonstrates missing data analysis using Pandas and Missingno.