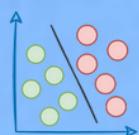
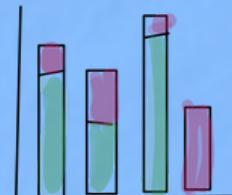
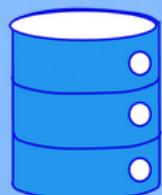
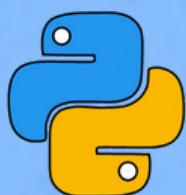


FREE

DATA SCIENCE

FULL ARCHIVE

250+ Python & Data
Science Posts



Daily Dose of
Data Science



DailyDoseofDS.com



Table of Contents

<i>Does Every ML Algorithm Rely on Gradient Descent?.....</i>	9
<i>Why Sklearn's Linear Regression Has No Hyperparameters?.....</i>	11
<i>Enrich The Default Preview of Pandas DataFrame with Jupyter DataTables.....</i>	13
<i>Visualize The Performance Of Linear Regression With This Simple Plot.....</i>	14
<i>Enrich Your Heatmaps With This Simple Trick.....</i>	16
<i>Confidence Interval and Prediction Interval Are Not The Same.....</i>	17
<i>The Ultimate Categorization of Performance Metrics in ML.....</i>	19
<i>The Coolest Matplotlib Hack to Create Subplots Intuitively.....</i>	23
<i>Execute Python Project Directory as a Script.....</i>	25
<i>The Most Overlooked Problem With One-Hot Encoding.....</i>	26
<i>9 Most Important Plots in Data Science.....</i>	28
<i>Is Categorical Feature Encoding Always Necessary Before Training ML Models?.</i>	30
<i>Scikit-LLM: Integrate Sklearn API with Large Language Models.....</i>	33
<i>The Counterintuitive Behaviour of Training Accuracy and Training Loss.....</i>	34
<i>A Highly Overlooked Point In The Implementation of Sigmoid Function.....</i>	38
<i>The Ultimate Categorization of Clustering Algorithms.....</i>	41
<i>Improve Python Run-time Without Changing A Single Line of Code.....</i>	43
<i>A Lesser-Known Feature of the Merge Method in Pandas.....</i>	45
<i>The Coolest GitHub-Colab Integration You Would Ever See.....</i>	47
<i>Most Sklearn Users Don't Know This About Its LinearRegression Implementation.....</i>	48
<i>Break the Linear Presentation of Notebooks With Stickyland.....</i>	50
<i>Visualize The Performance Of Any Linear Regression Model With This Simple Plot.....</i>	51
<i>Waterfall Charts: A Better Alternative to Line/Bar Plot.....</i>	53
<i>What Does The Google Styling Guide Say About Imports.....</i>	54
<i>How To Truly Use The Train, Validation and Test Set.....</i>	56
<i>Restart Jupyter Kernel Without Losing Variables.....</i>	59
<i>The Advantages and Disadvantages of PCA To Consider Before Using It.....</i>	60



<i>Loss Functions: An Algorithm-wise Comprehensive Summary.....</i>	62
<i>Is Data Normalization Always Necessary Before Training ML Models?.....</i>	64
<i>Annotate Data With The Click Of A Button Using Pigeon.....</i>	67
<i>Enrich Your Confusion Matrix With A Sankey Diagram.....</i>	68
<i>A Visual Guide to Stochastic, Mini-batch, and Batch Gradient Descent.....</i>	70
<i>A Lesser-Known Difference Between For-Loops and List Comprehensions.....</i>	73
<i>The Limitation of PCA Which Many Folks Often Ignore.....</i>	75
<i>Magic Methods: An Underrated Gem of Python OOP.....</i>	78
<i>The Taxonomy Of Regression Algorithms That Many Don't Bother To Remember</i>	
<i>81</i>	
<i>A Highly Overlooked Approach To Analysing Pandas DataFrames.....</i>	83
<i>Visualise The Change In Rank Over Time With Bump Charts.....</i>	84
<i>Use This Simple Technique To Never Struggle With TP, TN, FP and FN Again....</i>	85
<i>The Most Common Misconception About Inplace Operations in Pandas.....</i>	87
<i>Build Elegant Web Apps Right From Jupyter Notebook with Mercury.....</i>	89
<i>Become A Bilingual Data Scientist With These Pandas to SQL Translations.....</i>	91
<i>A Lesser-Known Feature of Sklearn To Train Models on Large Datasets.....</i>	93
<i>A Simple One-Liner to Create Professional Looking Matplotlib Plots.....</i>	95
<i>Avoid This Costly Mistake When Indexing A DataFrame.....</i>	97
<i>9 Command Line Flags To Run Python Scripts More Flexibly.....</i>	100
<i>Breathing KMeans: A Better and Faster Alternative to KMeans.....</i>	102
<i>How Many Dimensions Should You Reduce Your Data To When Using PCA?... </i>	105
<i>🚀 Mito Just Got Supercharged With AI!.....</i>	108
<i>Be Cautious Before Drawing Any Conclusions Using Summary Statistics.....</i>	110
<i>Use Custom Python Objects In A Boolean Context.....</i>	112
<i>A Visual Guide To Sampling Techniques in Machine Learning.....</i>	114
<i>You Were Probably Given Incomplete Info About A Tuple's Immutability.....</i>	118
<i>A Simple Trick That Significantly Improves The Quality of Matplotlib Plots....</i>	120
<i>A Visual and Overly Simplified Guide to PCA.....</i>	122
<i>Supercharge Your Jupyter Kernel With ipyflow.....</i>	125
<i>A Lesser-known Feature of Creating Plots with Plotly.....</i>	127
<i>The Limitation Of Euclidean Distance Which Many Often Ignore.....</i>	129
<i>Visualising The Impact Of Regularisation Parameter.....</i>	132



<i>AutoProfiler: Automatically Profile Your DataFrame As You Work.....</i>	134
<i>A Little Bit Of Extra Effort Can Hugely Transform Your Storytelling Skills.....</i>	136
<i>A Nasty Hidden Feature of Python That Many Programmers Aren't Aware Of....</i>	
<i>138</i>	
<i>Interactively Visualise A Decision Tree With A Sankey Diagram.....</i>	141
<i>Use Histograms With Caution. They Are Highly Misleading!.....</i>	143
<i>Three Simple Ways To (Instantly) Make Your Scatter Plots Clutter Free.....</i>	145
<i>A (Highly) Important Point to Consider Before You Use KMeans Next Time....</i>	148
<i>Why You Should Avoid Appending Rows To A DataFrame.....</i>	151
<i>Matplotlib Has Numerous Hidden Gems. Here's One of Them.....</i>	153
<i>A Counterintuitive Thing About Python Dictionaries.....</i>	155
<i>Probably The Fastest Way To Execute Your Python Code.....</i>	158
<i>Are You Sure You Are Using The Correct Pandas Terminologies?.....</i>	160
<i>Is Class Imbalance Always A Big Problem To Deal With?.....</i>	163
<i>A Simple Trick That Will Make Heatmaps More Elegant.....</i>	165
<i>A Visual Comparison Between Locality and Density-based Clustering.....</i>	167
<i>Why Don't We Call It Logistic Classification Instead?.....</i>	168
<i>A Typical Thing About Decision Trees Which Many Often Ignore.....</i>	170
<i>Always Validate Your Output Variable Before Using Linear Regression.....</i>	171
<i>A Counterintuitive Fact About Python Functions.....</i>	172
<i>Why Is It Important To Shuffle Your Dataset Before Training An ML Model....</i>	173
<i>The Limitations Of Heatmap That Are Slowing Down Your Data Analysis.....</i>	174
<i>The Limitation Of Pearson Correlation Which Many Often Ignore.....</i>	175
<i>Why Are We Typically Advised To Set Seeds for Random Generators?.....</i>	176
<i>An Underrated Technique To Improve Your Data Visualizations.....</i>	177
<i>A No-Code Tool to Create Charts and Pivot Tables in Jupyter.....</i>	178
<i>If You Are Not Able To Code A Vectorized Approach, Try This.....</i>	179
<i>Why Are We Typically Advised To Never Iterate Over A DataFrame?.....</i>	181
<i>Manipulating Mutable Objects In Python Can Get Confusing At Times.....</i>	182
<i>This Small Tweak Can Significantly Boost The Run-time of KMeans.....</i>	184
<i>Most Python Programmers Don't Know This About Python OOP.....</i>	186
<i>Who Said Matplotlib Cannot Create Interactive Plots?.....</i>	188
<i>Don't Create Messy Bar Plots. Instead, Try Bubble Charts!.....</i>	189



<i>You Can Add a List As a Dictionary's Key (Technically)!.....</i>	190
<i>Most ML Folks Often Neglect This While Using Linear Regression.....</i>	191
<i>35 Hidden Python Libraries That Are Absolute Gems.....</i>	192
<i>Use Box Plots With Caution! They May Be Misleading.....</i>	193
<i>An Underrated Technique To Create Better Data Plots.....</i>	194
<i>The Pandas DataFrame Extension Every Data Scientist Has Been Waiting For</i>	195
<i>Supercharge Shell With Python Using Xonsh.....</i>	196
<i>Most Command-line Users Don't Know This Cool Trick About Using Terminals....</i>	197
<i>A Simple Trick to Make The Most Out of Pivot Tables in Pandas.....</i>	198
<i>Why Python Does Not Offer True OOP Encapsulation.....</i>	199
<i>Never Worry About Parsing Errors Again While Reading CSV with Pandas....</i>	200
<i>An Interesting and Lesser-Known Way To Create Plots Using Pandas.....</i>	201
<i>Most Python Programmers Don't Know This About Python For-loops.....</i>	202
<i>How To Enable Function Overloading In Python.....</i>	203
<i>Generate Helpful Hints As You Write Your Pandas Code.....</i>	204
<i>Speedup NumPy Methods 25x With Bottleneck.....</i>	205
<i>Visualizing The Data Transformation of a Neural Network.....</i>	206
<i>Never Refactor Your Code Manually Again. Instead, Use Sourcery!.....</i>	207
<i>Draw The Data You Are Looking For In Seconds.....</i>	208
<i>Style Matplotlib Plots To Make Them More Attractive.....</i>	209
<i>Speed-up Parquet I/O of Pandas by 5x.....</i>	210
<i>40 Open-Source Tools to Supercharge Your Pandas Workflow.....</i>	211
<i>Stop Using The Describe Method in Pandas. Instead, use Skimpy.....</i>	212
<i>The Right Way to Roll Out Library Updates in Python.....</i>	213
<i>Simple One-Liners to Preview a Decision Tree Using Sklearn.....</i>	214
<i>Stop Using The Describe Method in Pandas. Instead, use Summarytools.....</i>	215
<i>Never Search Jupyter Notebooks Manually Again To Find Your Code.....</i>	216
<i>F-strings Are Much More Versatile Than You Think.....</i>	217
<i>Is This The Best Animated Guide To KMeans Ever?.....</i>	218
<i>An Effective Yet Underrated Technique To Improve Model Performance.....</i>	219
<i>Create Data Plots Right From The Terminal.....</i>	220
<i>Make Your Matplotlib Plots More Professional.....</i>	221



37 Hidden Python Libraries That Are Absolute Gems.....	222
Preview Your README File Locally In GitHub Style.....	223
Pandas and NumPy Return Different Values for Standard Deviation. Why?... 	224
Visualize Commit History of Git Repo With Beautiful Animations.....	225
Perfplot: Measure, Visualize and Compare Run-time With Ease.....	226
This GUI Tool Can Possibly Save You Hours Of Manual Work.....	227
How Would You Identify Fuzzy Duplicates In A Data With Million Records?....	228
Stop Previewing Raw DataFrames. Instead, Use DataTables.....	230
🚀 A Single Line That Will Make Your Python Code Faster.....	231
Prettify Word Clouds In Python.....	232
How to Encode Categorical Features With Many Categories?.....	233
Calendar Map As A Richer Alternative to Line Plot.....	234
10 Automated EDA Tools That Will Save You Hours Of (Tedious) Work.....	235
Why KMeans May Not Be The Apt Clustering Algorithm Always.....	236
Converting Python To LaTeX Has Possibly Never Been So Simple.....	237
Density Plot As A Richer Alternative to Scatter Plot.....	238
30 Python Libraries to (Hugely) Boost Your Data Science Productivity.....	239
Sklearn One-liner to Generate Synthetic Data.....	240
Label Your Data With The Click Of A Button.....	241
Analyze A Pandas DataFrame Without Code.....	242
Python One-Liner To Create Sketchy Hand-drawn Plots.....	243
70x Faster Pandas By Changing Just One Line of Code.....	244
An Interactive Guide To Master Pandas In One Go.....	245
Make Dot Notation More Powerful in Python.....	246
The Coolest Jupyter Notebook Hack.....	247
Create a Moving Bubbles Chart in Python.....	248
Skorch: Use Scikit-learn API on PyTorch Models.....	249
Reduce Memory Usage Of A Pandas DataFrame By 90%.....	250
An Elegant Way To Perform Shutdown Tasks in Python.....	251
Visualizing Google Search Trends of 2022 using Python.....	252
Create A Racing Bar Chart In Python.....	253
Speed-up Pandas Apply 5x with NumPy.....	254



<i>A No-Code Online Tool To Explore and Understand Neural Networks.....</i>	255
<i>What Are Class Methods and When To Use Them?.....</i>	256
<i>Make Sklearn KMeans 20x times faster.....</i>	257
<i>Speed-up NumPy 20x with Numexpr.....</i>	258
<i>A Lesser-Known Feature of Apply Method In Pandas.....</i>	259
<i>An Elegant Way To Perform Matrix Multiplication.....</i>	260
<i>Create Pandas DataFrame from Dataclass.....</i>	261
<i>Hide Attributes While Printing A Dataclass Object.....</i>	262
<i>List : Tuple :: Set : ?.....</i>	263
<i>Difference Between Dot and Matmul in NumPy.....</i>	264
<i>Run SQL in Jupyter To Analyze A Pandas DataFrame.....</i>	265
<i>Automated Code Refactoring With Sourcery.....</i>	266
<i>__Post_init__: Add Attributes To A Dataclass Object Post Initialization.....</i>	267
<i>Simplify Your Functions With Partial Functions.....</i>	268
<i>When You Should Not Use the head() Method In Pandas.....</i>	269
<i>DotMap: A Better Alternative to Python Dictionary.....</i>	270
<i>Prevent Wild Imports With __all__ in Python.....</i>	271
<i>Three Lesser-known Tips For Reading a CSV File Using Pandas.....</i>	272
<i>The Best File Format To Store A Pandas DataFrame.....</i>	273
<i>Debugging Made Easy With PySnooper.....</i>	274
<i>Lesser-Known Feature of the Merge Method in Pandas.....</i>	275
<i>The Best Way to Use Apply() in Pandas.....</i>	276
<i>Deep Learning Network Debugging Made Easy.....</i>	277
<i>Don't Print NumPy Arrays! Use Lovely-NumPy Instead.....</i>	278
<i>Performance Comparison of Python 3.11 and Python 3.10.....</i>	279
<i>View Documentation in Jupyter Notebook.....</i>	280
<i>A No-code Tool To Understand Your Data Quickly.....</i>	281
<i>Why 256 is 256 But 257 is not 257?.....</i>	282
<i>Make a Class Object Behave Like a Function.....</i>	284
<i>Lesser-known feature of Pickle Files.....</i>	286
<i>Dot Plot: A Potential Alternative to Bar Plot.....</i>	288
<i>Why Correlation (and Other Statistics) Can Be Misleading.....</i>	289



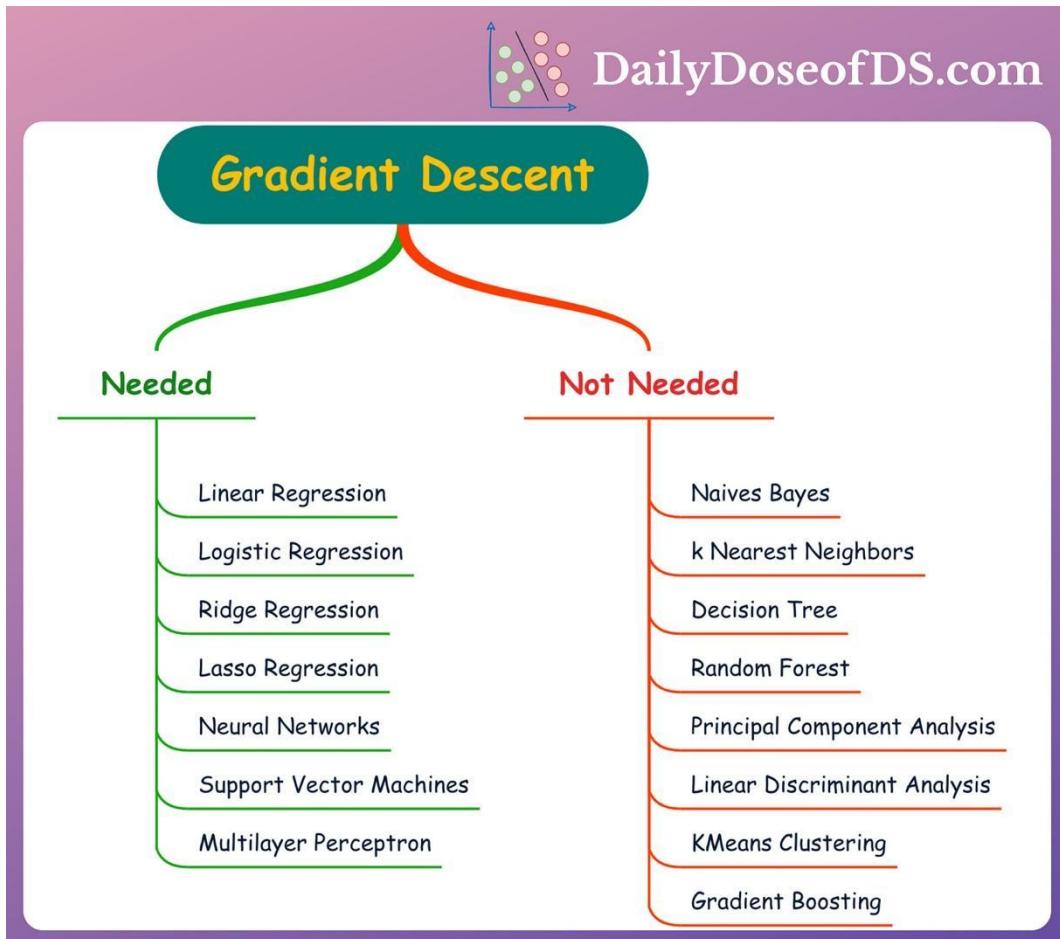
<i>Supercharge value_counts() Method in Pandas With Sidetable.....</i>	290
<i>Write Your Own Flavor Of Pandas.....</i>	291
<i>CodeSquire: The AI Coding Assistant You Should Use Over GitHub Copilot.....</i>	292
<i>Vectorization Does Not Always Guarantee Better Performance.....</i>	293
<i>In Defense of Match-case Statements in Python.....</i>	294
<i>Enrich Your Notebook With Interactive Controls.....</i>	296
<i>Get Notified When Jupyter Cell Has Executed.....</i>	298
<i>Data Analysis Using No-Code Pandas In Jupyter.....</i>	299
<i>Using Dictionaries In Place of If-conditions.....</i>	300
<i>Clear Cell Output In Jupyter Notebook During Run-time.....</i>	302
<i>A Hidden Feature of Describe Method In Pandas.....</i>	303
<i>Use Slotted Class To Improve Your Python Code.....</i>	304
<i>Stop Analysing Raw Tables. Use Styling Instead!.....</i>	305
<i>Explore CSV Data Right From The Terminal.....</i>	306
<i>Generate Your Own Fake Data In Seconds.....</i>	307
<i>Import Your Python Package as a Module.....</i>	308
<i>Specify Loops and Runs In %%timeit.....</i>	309
<i>Waterfall Charts: A Better Alternative to Line/Bar Plot.....</i>	310
<i>Hexbin Plots As A Richer Alternative to Scatter Plots.....</i>	311
<i>Importing Modules Made Easy with Pyforest.....</i>	312
<i>Analyse Flow Data With Sankey Diagrams.....</i>	314
<i>Feature Tracking Made Simple In Sklearn Transformers.....</i>	316
<i>Lesser-known Feature of f-strings in Python.....</i>	318
<i>Don't Use time.time() To Measure Execution Time.....</i>	319
<i>Now You Can Use DALL-E With OpenAI API.....</i>	320
<i>Polynomial Linear Regression Plot Made Easy With Seaborn.....</i>	321
<i>Retrieve Previously Computed Output In Jupyter Notebook.....</i>	322
<i>Parallelize Pandas Apply() With Swifter.....</i>	323
<i>Create DataFrame Hassle-free By Using Clipboard.....</i>	324
<i>Run Python Project Directory As A Script.....</i>	325
<i>Inspect Program Flow with IceCream.....</i>	326
<i>Don't Create Conditional Columns in Pandas with Apply.....</i>	327



<i>Pretty Plotting With Pandas.....</i>	328
<i>Build Baseline Models Effortlessly With Sklearn.....</i>	329
<i>Fine-grained Error Tracking With Python 3.11.....</i>	330
<i>Find Your Code Hiding In Some Jupyter Notebook With Ease.....</i>	331
<i>Restart the Kernel Without Losing Variables.....</i>	332
<i>How to Read Multiple CSV Files Efficiently.....</i>	333
<i>Elegantly Plot the Decision Boundary of a Classifier.....</i>	335
<i>An Elegant Way to Import Metrics From Sklearn.....</i>	336
<i>Configure Sklearn To Output Pandas DataFrame.....</i>	337
<i>Display Progress Bar With Apply() in Pandas.....</i>	338
<i>Modify a Function During Run-time.....</i>	339
<i>Regression Plot Made Easy with Plotly.....</i>	340
<i>Polynomial Linear Regression with NumPy.....</i>	341
<i>Alter the Datatype of Multiple Columns at Once.....</i>	342
<i>Datatype For Handling Missing Valued Columns in Pandas.....</i>	343
<i>Parallelize Pandas with Pandarallel.....</i>	344
<i>Why you should not dump DataFrames to a CSV.....</i>	345
<i>Save Memory with Python Generators.....</i>	347
<i>Don't use print() to debug your code.....</i>	348
<i>Find Unused Python Code With Ease.....</i>	350
<i>Define the Correct DataType for Categorical Columns.....</i>	351
<i>Transfer Variables Between Jupyter Notebooks.....</i>	352
<i>Why You Should Not Read CSVs with Pandas.....</i>	353
<i>Modify Python Code During Run-Time.....</i>	354
<i>Handle Missing Data With Missingno.....</i>	355



Does Every ML Algorithm Rely on Gradient Descent?



Gradient descent is the most common optimization technique in ML. Essentially, the core idea is to iteratively update the model's parameters by calculating the gradients of the cost function with respect to those parameters.

Why gradient descent is a critical technique, it is important to know that not all algorithms rely on gradient descent.

The visual above depicts this.

Algorithms that rely on gradient descent:

- Linear Regression
- Logistic Regression
- Ridge Regression
- Lasso Regression
- Neural Networks (ANNs, RNNs, CNNs, LSTMs, etc.)



- Support Vector Machines
- Multilayer Perceptrons

Algorithms that DON'T rely on gradient descent:

- Naive Bayes
- kNN
- Decision Tree
- Random Forest
- Principal Component Analysis
- Linear Discriminant Analysis
- KMeans Clustering
- Gradient Boosting



Why Sklearn's Linear Regression Has No Hyperparameters?

sklearn.linear_model.LinearRegression

```
class sklearn.linear_model.LinearRegression(*, fit_intercept=True, copy_X=True, n_jobs=None, positive=False)
```

[source]

No Hyperparameters!

Parameters:

- fit_intercept : bool, default=True**
Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations (i.e. data is expected to be centered).
- copy_X : bool, default=True**
If True, X will be copied; else, it may be overwritten.
- n_jobs : int, default=None**
The number of jobs to use for the computation. This will only provide speedup in case of sufficiently large problems, that is if firstly `n_targets > 1` and secondly `X` is sparse or if `positive` is set to `True`. `None` means 1 unless in a `joblib.parallel_backend` context. `-1` means using all processors. See `Glossary` for more details.
- positive : bool, default=False**
When set to `True`, forces the coefficients to be positive. This option is only supported for dense arrays.

New in version 0.24.

All ML models we work with have some hyperparameters, such as:

- Learning rate
- Regularization
- Layer size (for neural network), etc.

But as shown in the image above, why don't we see one in Sklearn's Linear Regression implementation?

To understand the reason, we first need to realize that the Linear Regression algorithm can model data in two different ways:

Gradient Descent (which we use with almost all other ML algorithms):

- It is a stochastic algorithm, i.e., involves some randomness.
- It finds an approximate solution using optimization.
- It has hyperparameters.



Ordinary Least Square (OLS):

- It is a deterministic algorithm. If run multiple times, it will always converge to the same weights.
- It always finds the optimal solution.
- It has no hyperparameters.

Instead of gradient descent, Sklearn's Linear Regression class implements the OLS method.

sklearn.linear_model.LinearRegression

```
class sklearn.linear_model.LinearRegression(*, fit_intercept=True, copy_X=True, n_jobs=None,
                                         positive=False)
```

[source]

Ordinary least squares Linear Regression.

OLS

That is why it has no hyperparameters.

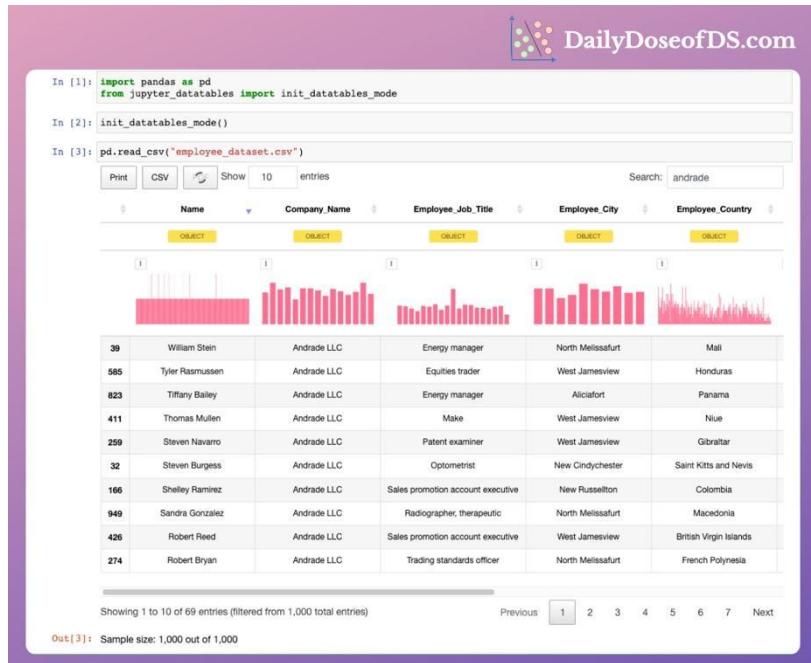
How does OLS work?

Read the full post here with equations:

<https://www.blog.dailydoseofds.com/p/why-sklearns-linear-regression-has>.



Enrich The Default Preview of Pandas DataFrame with Jupyter DataTables



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.

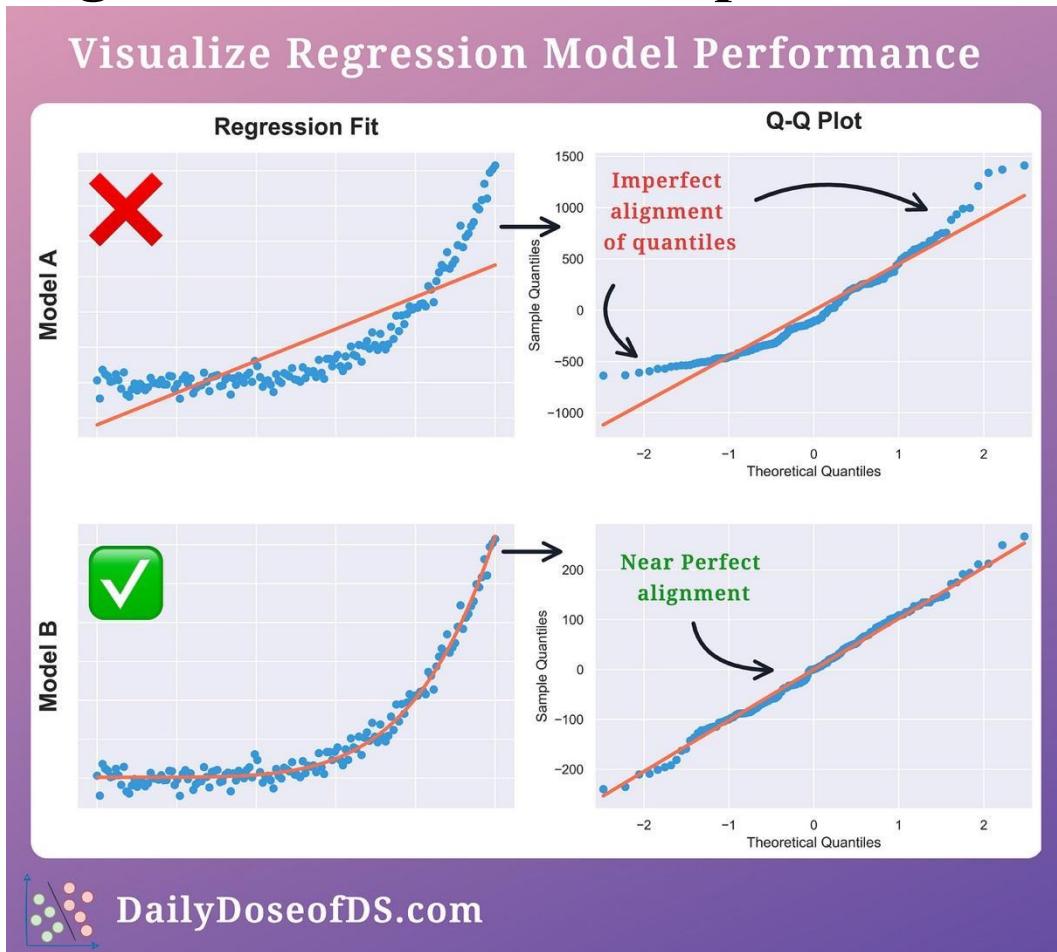
This includes:

- sorting
- filtering
- exporting
- plotting column distribution
- printing data types,
- pagination, and more.

Check it out here: [GitHub](#).



Visualize The Performance Of Linear Regression With This Simple Plot



Linear regression assumes that the model residuals (=actual-predicted) are normally distributed.

If the model is underperforming, it may be due to a violation of this assumption.

A QQ plot (short for Quantile-Quantile) is a great way to verify this and also determine the model's performance.

As the name suggests, it depicts the quantiles of the observed distribution (residuals in this case) against the quantiles of a reference distribution, typically the standard normal distribution.

A good QQ plot will:

- Show minimal deviations from the reference line, indicating that the residuals are approximately normally distributed.

A bad QQ plot will:



- Exhibit significant deviations, indicating a departure from the normality of residuals.
- Display patterns of skewness with its diverging ends, etc.

Thus, the more aligned the QQ plot looks, the more confident you can be about your model.

This is especially useful when the regression line is difficult to visualize, i.e., in a high-dimensional dataset.

So remember...

After running a linear model, always check the distribution of the residuals.

This will help you:

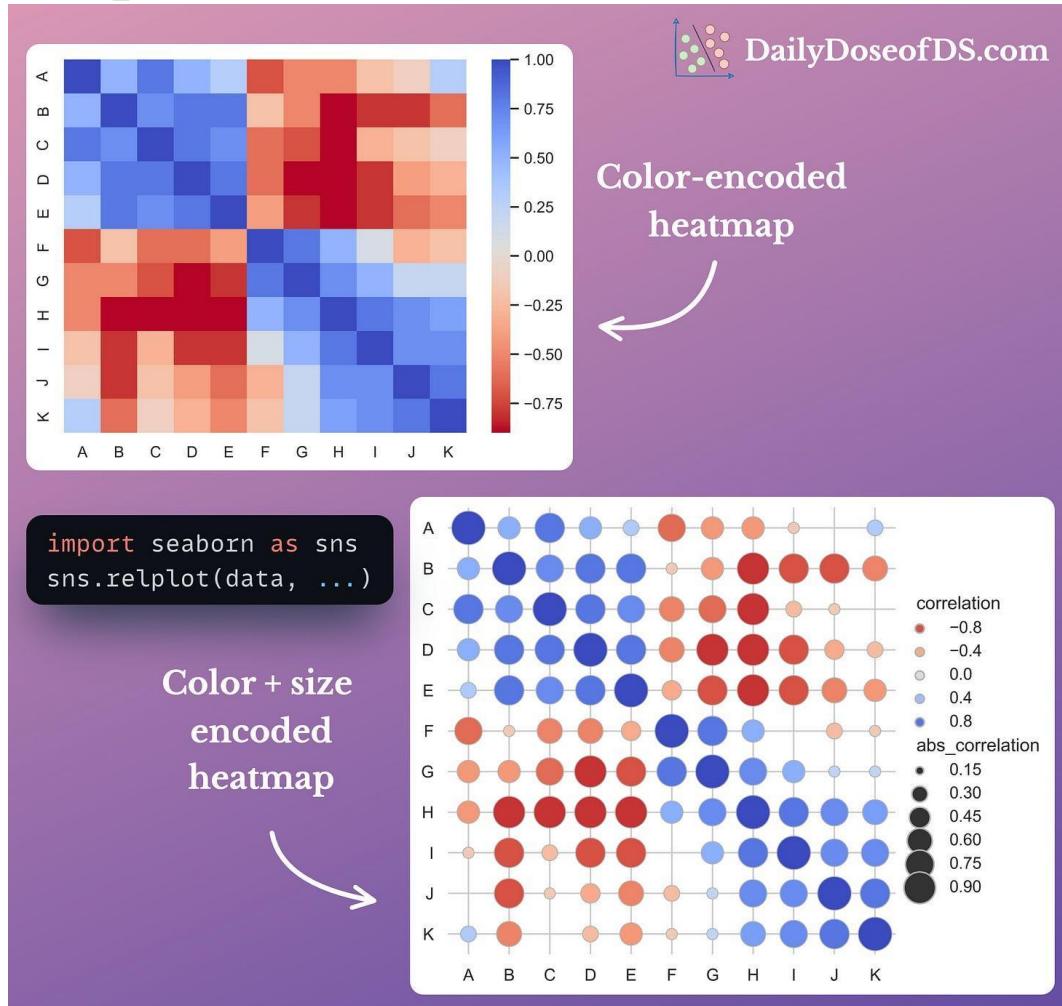
- Validate the model's assumptions
- Determine how good your model is
- Find ways to improve it (if needed)

👉 Over to you: What are some other ways/plots to determine the linear model's performance?

I covered another way in one of my previous posts: [Visualize The Performance Of Any Linear Regression Model With This Simple Plot.](#)



Enrich Your Heatmaps With This Simple Trick



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.

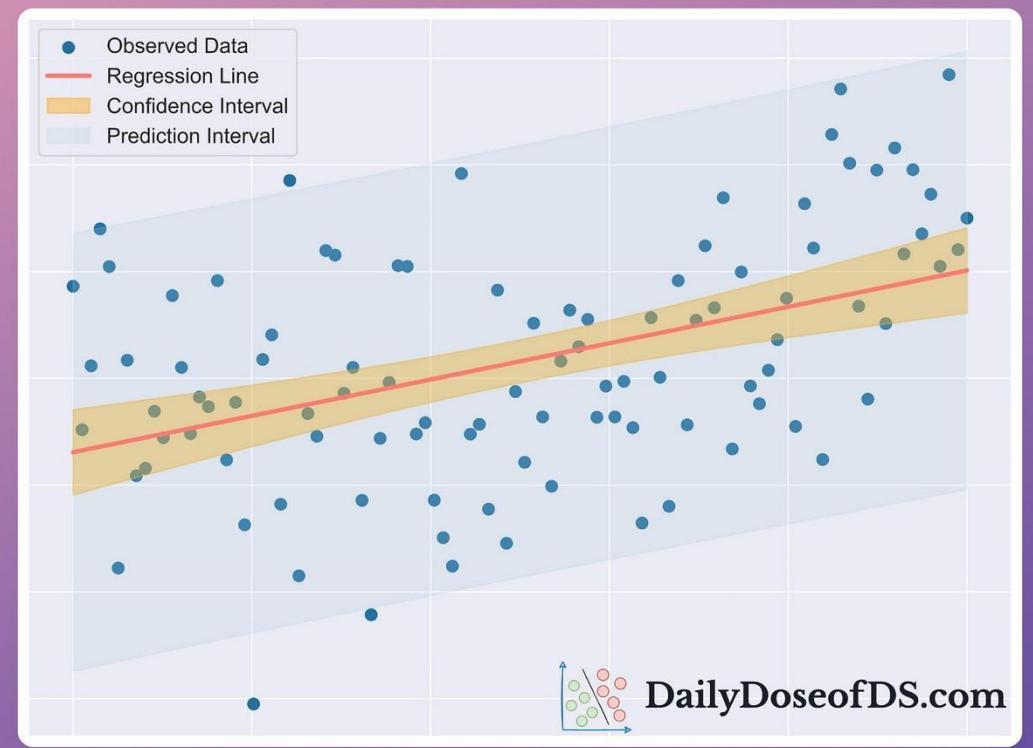
This is especially useful to make heatmaps cleaner, as many values nearer to zero will immediately shrink.

Find more details here: [Seaborn Docs](https://seaborn.pydata.org/).



Confidence Interval and Prediction Interval Are Not The Same

Confidence Interval & Prediction Interval



Contrary to common belief, linear regression NEVER predicts an actual value.

Instead, it models the relationship between the input and an average related to the outcome.

Thus, there's always some uncertainty involved, and it is important to communicate it.

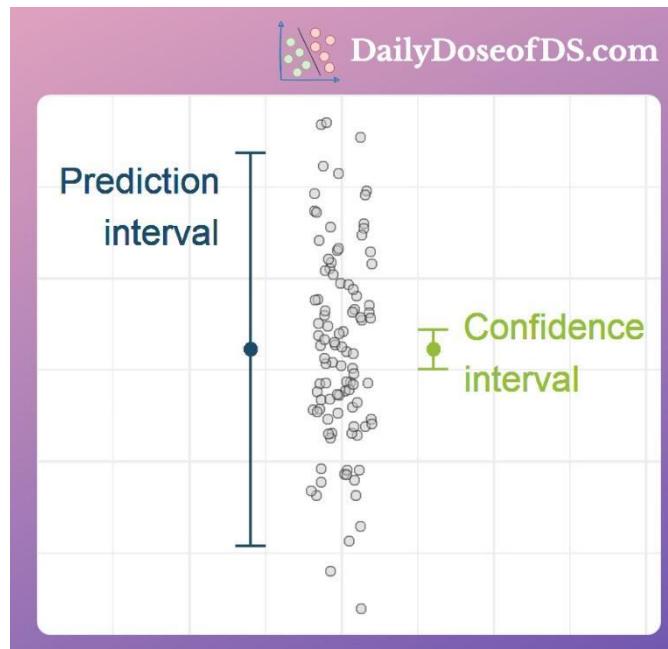
Confidence interval and prediction interval help us capture this uncertainty.

Confidence interval:

- tells the range of mean outcome at a given input.
- answers the question: "If we know the input, what is the uncertainty around the average value of the outcome."

Thus, a 95% confidence interval says:

- given an input, we are 95% confident that the actual mean will lie in that region.



Prediction interval, however:

- tells the range of possible values the outcome variable may take.
- answers the question: "If we know the input, what is the actual range of the outcome variable that we may observe."

For instance, a 95% prediction interval tells us that:

- given an input, 95% of observed values will lie in that region.

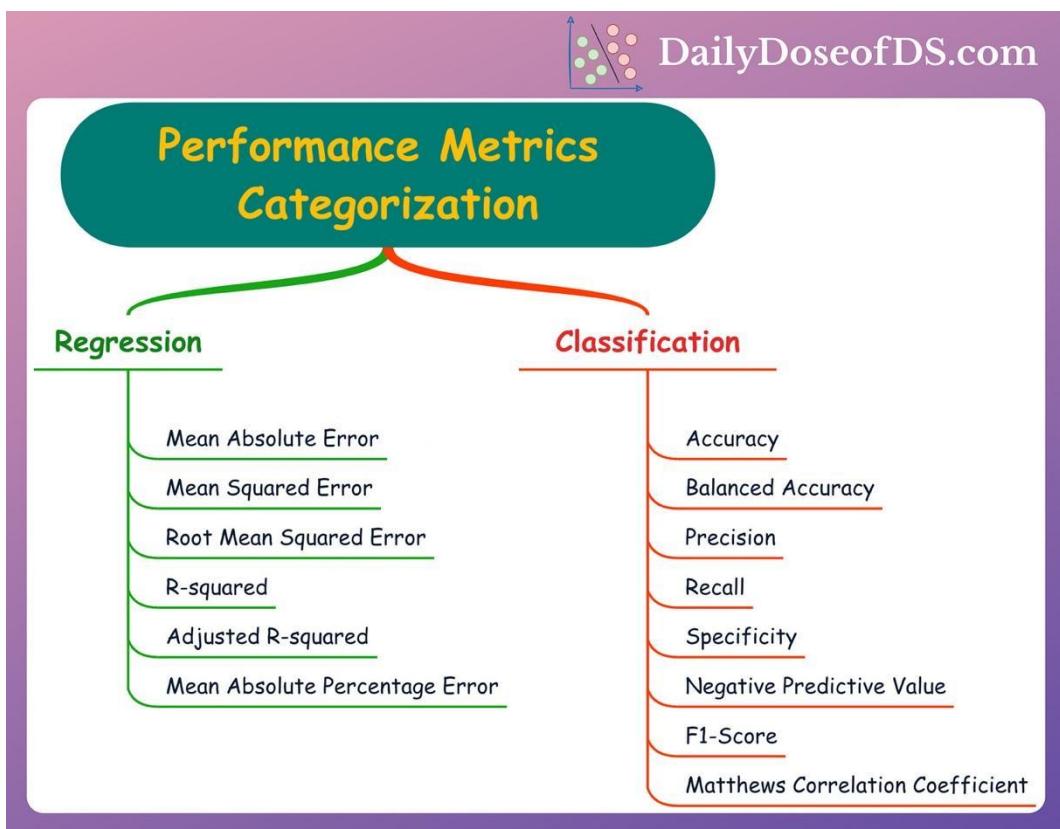
So remember...

- Confidence interval and prediction interval are NOT the same.
- They depict different uncertainties of the outcome variable.
- Confidence interval captures the range of the mean outcome around an input.
- Prediction interval captures the range of the actual values of the outcome around an input.
- Prediction interval is typically wider than confidence interval.

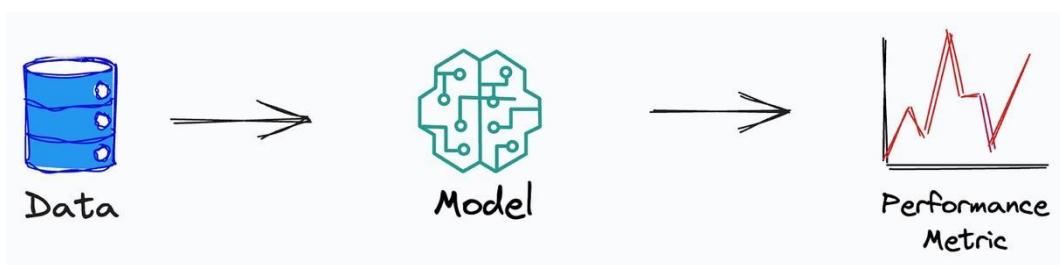
👉 Over to you: What does a 100% confidence interval and prediction interval will look like?



The Ultimate Categorization of Performance Metrics in ML



Performance metrics are used to assess the performance of a model on a given dataset.



They provide quantitative ways to test the effectiveness of a model. They also help in identifying the areas for improvement and optimization.

Why Performance metrics?

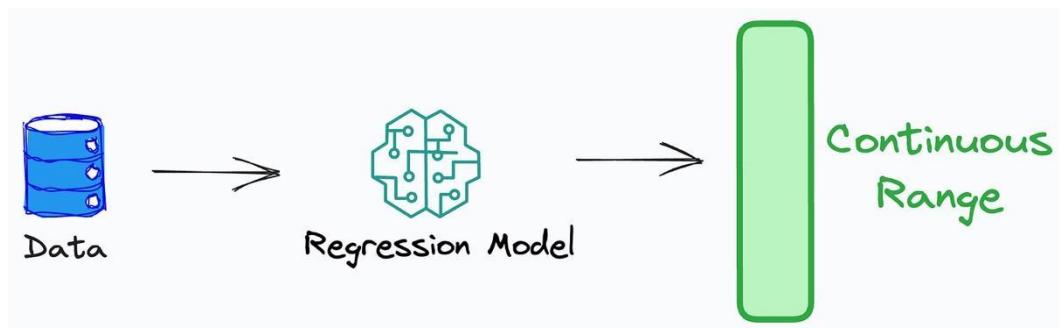


Typically, it is difficult to interpret ML models (especially deep learning models). It is difficult to understand the specific patterns identified by the model from the given data.

Performance metrics allow us to determine their performance by providing unseen samples by evaluating the predictions.

This makes them a must-know in ML.

Performance Metrics for Regression



Mean Absolute Error (MAE):

Measures the average absolute difference between the predicted and actual value.

Provides a straightforward assessment of prediction accuracy.

Mean Squared Error (MSE):

Measures the average squared difference between the predicted and actual values.

Larger errors inflate the overall metric.

Root Mean Squared Error:

It is the square root of MSE.

R-squared:

Represents the proportion of the variance in the target variable explained by the regression model.

Adjusted R-squared:

Similar to R-squared.

But accounts for the number of predictors (features) in the model.

Penalizes model complexity.

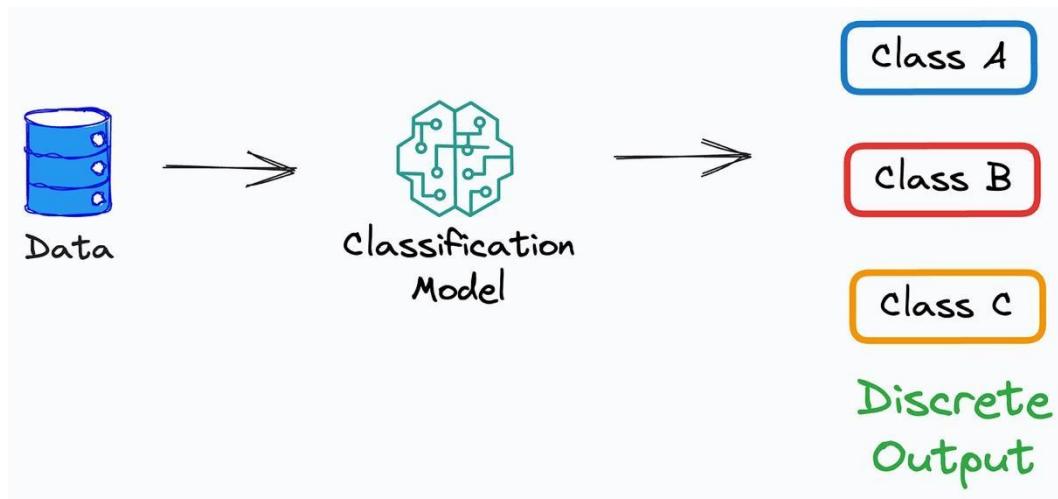


Mean Absolute Percentage Error (MAPE):

Measures the average percentage difference between the predicted and actual values.

Typically used when the scale of the target variable is significant.

Performance Metrics for Classification



Accuracy:

Measures the proportion of correct predictions, irrespective of the class.

Provides an overall assessment of classification performance.

Precision:

Measures the proportion of positive predictions that were correct.

It is also called the accuracy of the model only on the positive predictions.

Recall (Sensitivity):

Measures the proportion of correctly classified positive samples in the dataset.

It is also called the accuracy of the model on positive instances in the dataset.

Precision → Accuracy on positive predictions.

Recall → Accuracy on positive instances in the dataset.

Read by Medium blog to understand more: [Precision-Recall Blog](#).



Specificity:

Opposite of Recall.

Measures the proportion of correctly classified negative samples in the dataset.

It is also called the accuracy of the model on negative instances in the dataset.

Negative Predictive Value:

Opposite of Precision.

Measures the proportion of negative predictions that were correct.

It is also called the accuracy of the model only on the negative predictions.

Balanced Accuracy:

Computes the average of Recall (accuracy on positive predictions) and specificity (accuracy on negative predictions)

A better and less-misleading measure than Accuracy in the case of an imbalanced dataset.

F1-score:

The harmonic mean of precision and recall.

Provides a balanced measure between the two.

Matthews Correlation Coefficient (MCC):

Takes into account true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions to measure the performance of the binary classifier.

Provides a balanced performance measure unaffected by class imbalance.

If you struggle to understand TP, TN, FP and FN, read my previous post: [Use This Simple Technique To Never Struggle With TP, TN, FP and FN Again](#)



The Coolest Matplotlib Hack to Create Subplots Intuitively



This has to be the coolest thing I have ever learned about Matplotlib.

We mostly use `plt.subplots()` method to create subplots using Matplotlib.

But this, at times, gets pretty tedious and cumbersome. For instance:

it offers limited flexibility to create a custom layout.

it is prone to indexing errors, and more.

Instead, use the `plt.subplot_mosaic()` method.

Here, you can create a plot of any desired layout by defining the plot structure as a string.

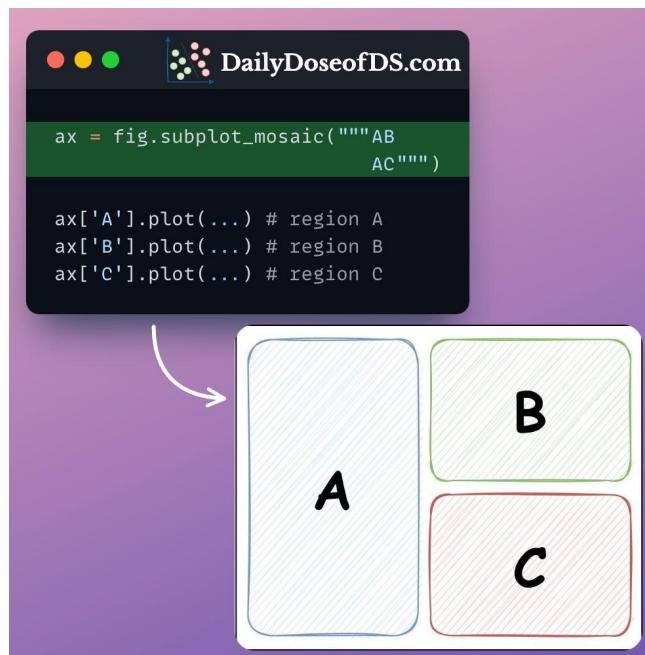
For instance, the string layout:



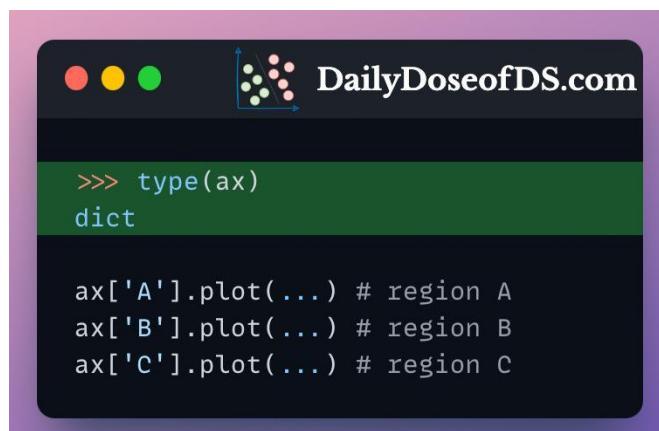
- AB
- AC

will create three subplots, wherein:

- subplot "A" will span the full first column
- subplot "B" will span the top half of the second column
- subplot "C" will span the bottom half of the second column



Next, create a subplot in a specific region by indexing the axes dictionary with its subplot key ("A", "B", or "C").

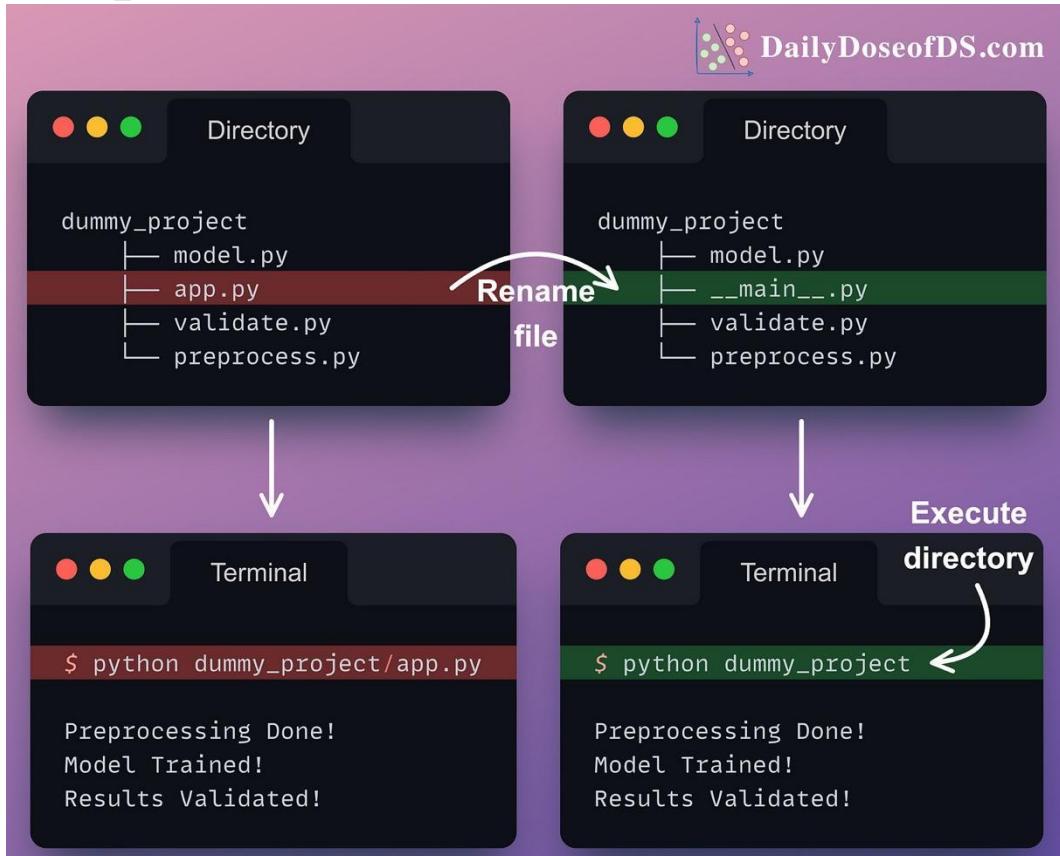


Isn't that super convenient and cool?

Read more: [Matplotlib Docs](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplot_mosaic.html).



Execute Python Project Directory as a Script



We mostly run a Python pipeline by invoking a script (**.py** file).

But did you know you can also execute the Python project directory as a script?

To do this, rename the base file of your project to **__main__.py**.

As a result, you can execute the whole pipeline by running the parent directory itself.

This is concise and elegant.

It also makes it easier for other users to use your project, as they are not required to dig into the directory and locate the base file.



The Most Overlooked Problem With One-Hot Encoding

The correct way to one-hot encode data

DailyDoseofDS.com

Categorical data

size	
0	small
1	medium
2	large

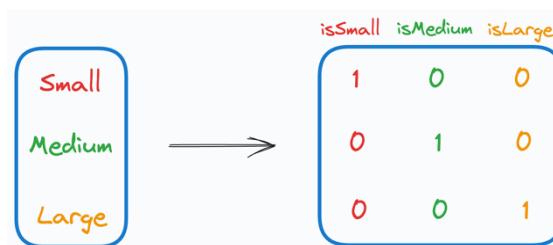
Don't keep all features

	size_large	size_medium	size_small
0	0	0	1
1	0	1	0
2	1	0	0

Instead, drop one feature

	size_medium	size_small
0	0	1
1	1	0
2	0	0

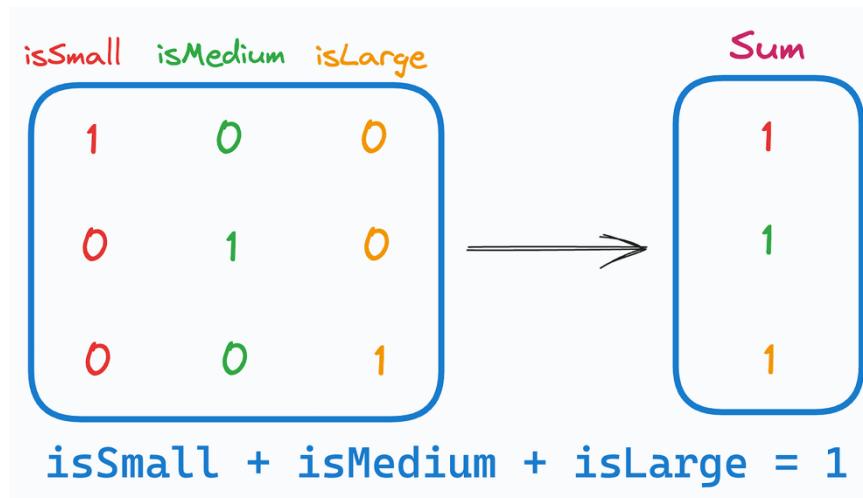
With one-hot encoding, we introduce a big problem in the data.



When we one-hot encode categorical data, we unknowingly introduce perfect multicollinearity.

Multicollinearity arises when two or more features can predict another feature.

As the sum of one-hot encoded features is always 1, it leads to perfect multicollinearity.



This is often called the Dummy Variable Trap.

It is bad because:

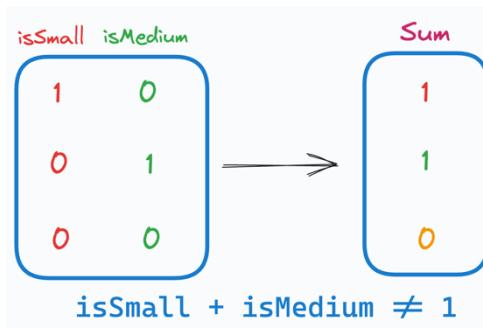
- The model has redundant features
- Regressions coefficients aren't reliable in the presence of multicollinearity, etc.

So how to resolve this?

The solution is simple.

Drop any arbitrary feature from the one-hot encoded features.

This instantly mitigates multicollinearity and breaks the linear relationship which existed before.



So remember...

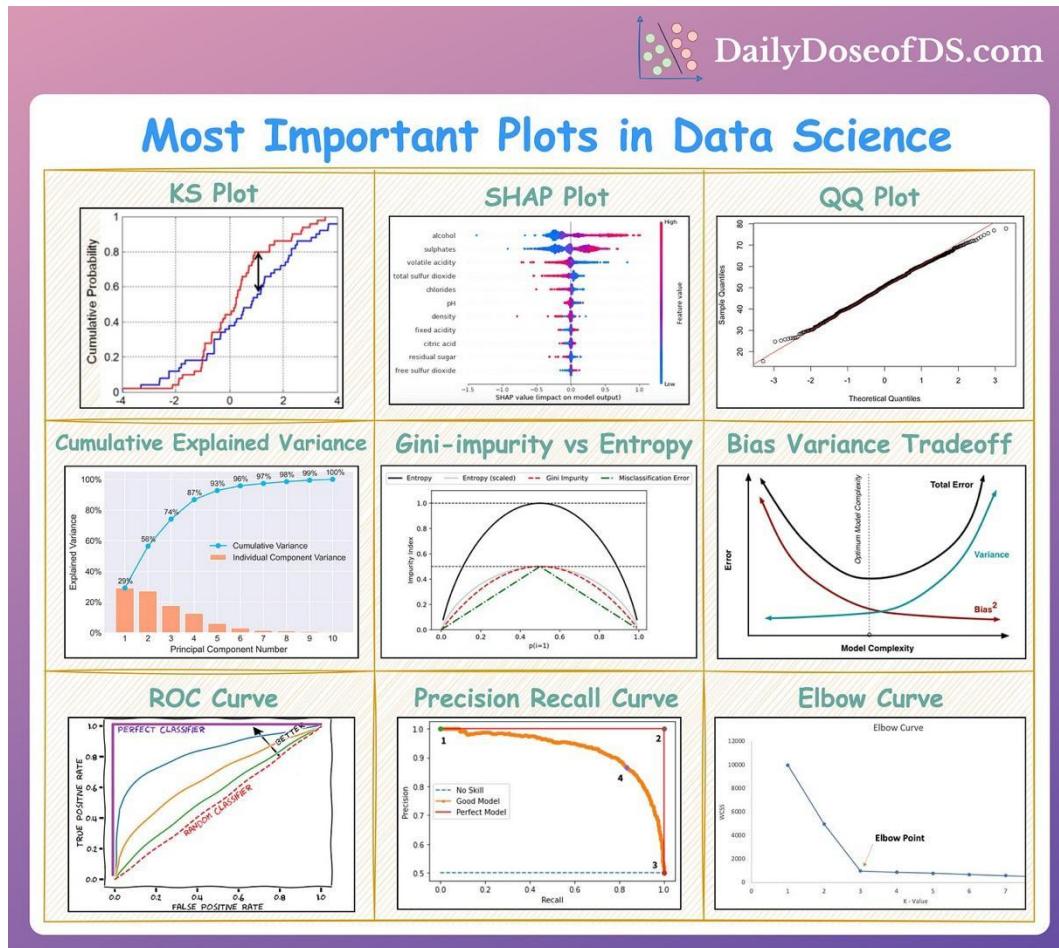
Whenever we one-hot encode categorical data, it introduces multicollinearity.

To avoid this, drop one column and proceed ahead.

👉 Over to you: What are some other problems with one-hot encoding?



9 Most Important Plots in Data Science



Exploring and analyzing data is a fundamental aspect of data science.

Here, visualizations play a crucial role in understanding complex patterns and relationships.

They offer a concise way to:

- understand the intricacies of statistical models,
- validate model assumptions,
- evaluate model performance, and much more.

The visual above depicts 9 of the most important and must-know plots in data science.

- **KS Plot:** It compares the cumulative distribution functions (CDFs) of a dataset to a theoretical distribution or between two datasets to assess the distributional differences.

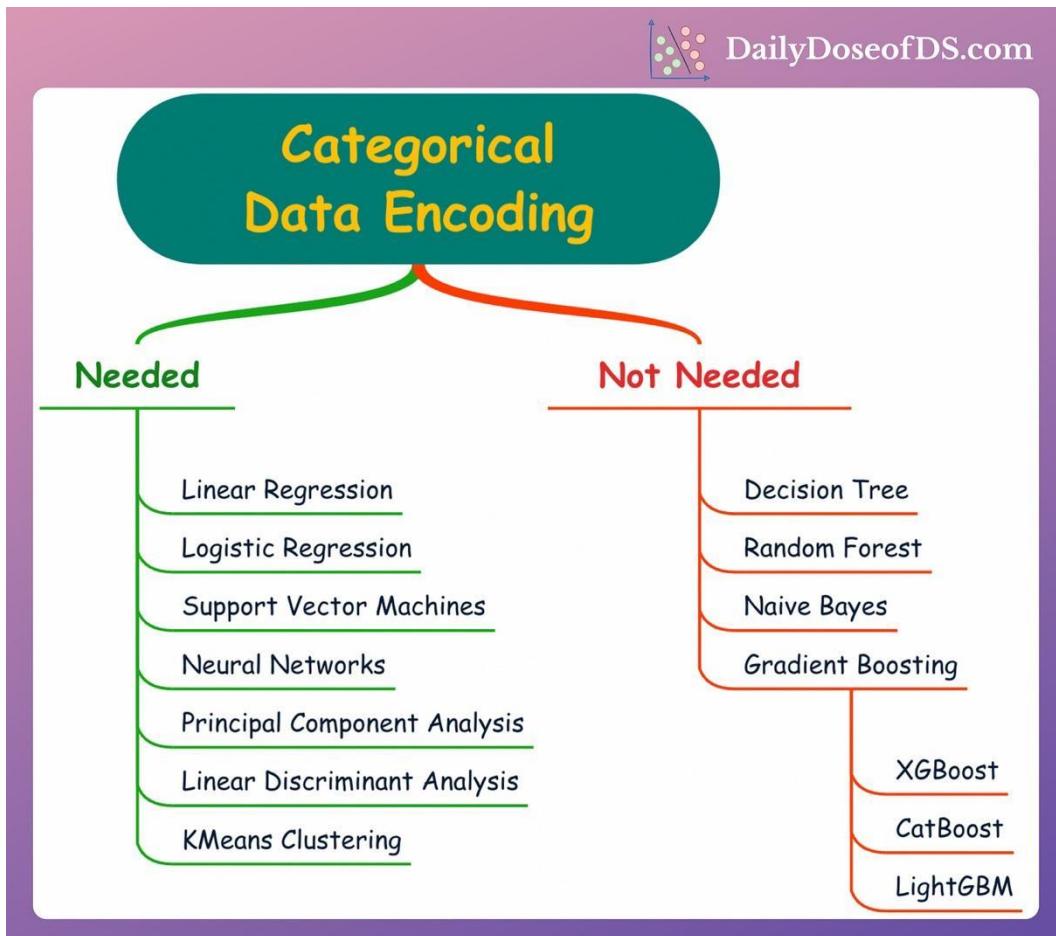


- **SHAP Plot:** It provides a summary of feature importance to a model's predictions, by considering interactions/dependencies between them.
- **QQ Plot:** It is used to assess the distributional similarity between observed data and theoretical distribution.
 - Here, we plot the quantiles of the two distributions against each other.
 - Deviations from the straight line indicate a departure from the assumed distribution.
- **Cumulative Explained Variance Plot:** I covered this in a detailed post before: [How Many Dimensions Should You Reduce Your Data To When Using PCA?](#)
- **Gini-Impurity vs. Entropy:** They are used to measure the impurity or disorder of a node or split in a decision tree.
- The plot compares Gini impurity and Entropy across different splits. This provides insights into the tradeoff between these measures.
- **Bias-Variance Tradeoff:** It is used to find the right balance between the bias and the variance of a model.
- **ROC Curve:** It depicts the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different classification thresholds.
- **Precision-Recall Curve:** It depicts the trade-off between Precision and Recall across different classification thresholds.
- **Elbow Curve:** The plot helps identify the optimal number of clusters for k-means algorithm.

Over to you: What more plots will you include here?



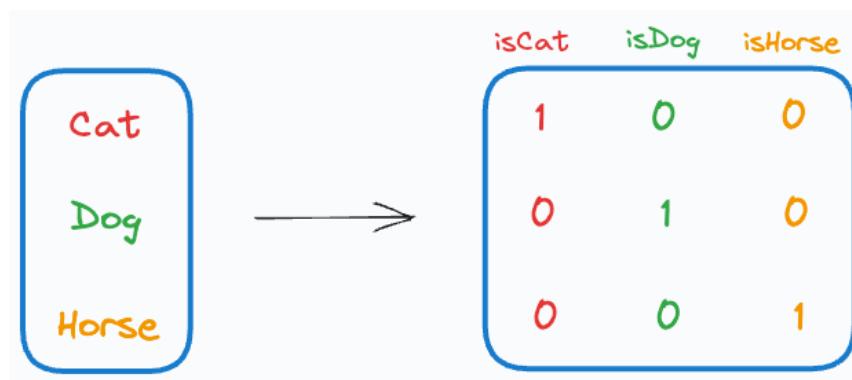
Is Categorical Feature Encoding Always Necessary Before Training ML Models?



When data contains categorical features, they may need special attention at times. This is because many algorithms require numerical data to work with.

Thus, when dealing with such datasets, it becomes crucial to handle these features appropriately to ensure accurate and meaningful analysis.

For instance, one common approach is to use one-hot encoding, as shown below:

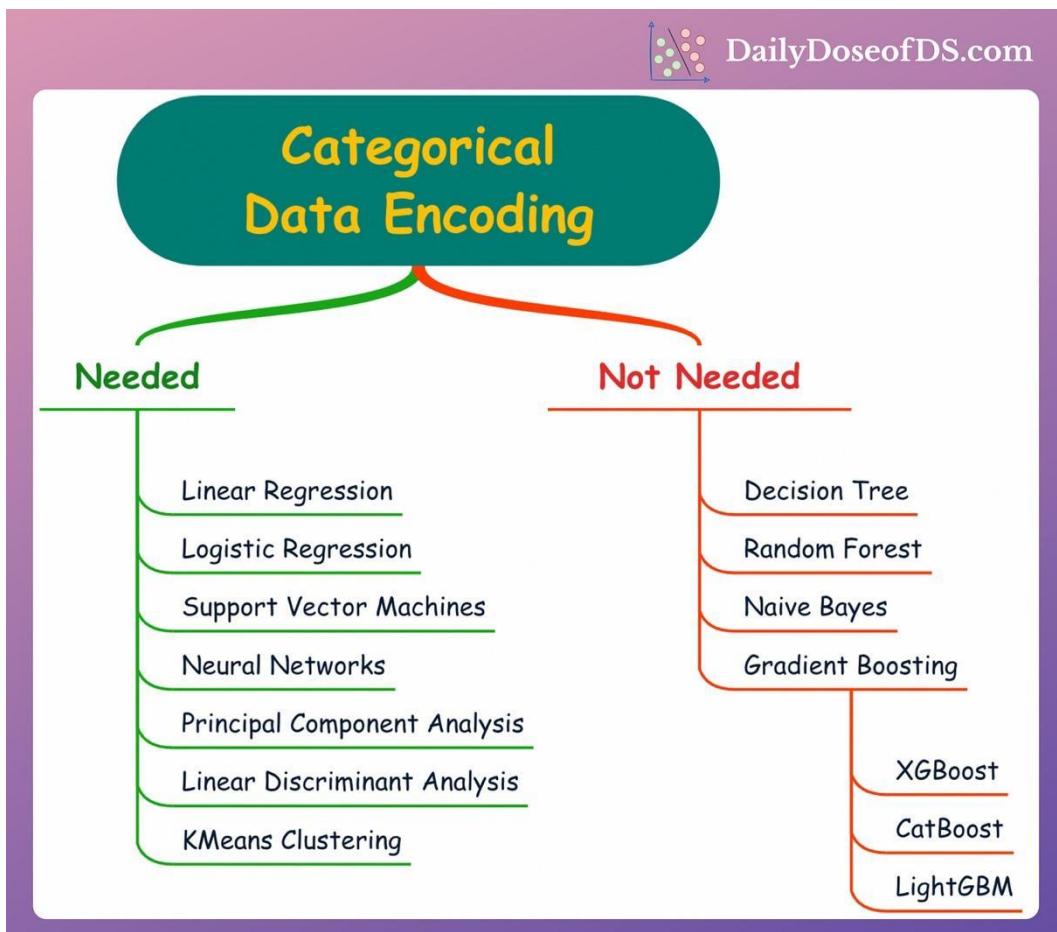


Encoding categorical data allows algorithms to process them effectively.

But is it always necessary?

While encoding categorical data is often crucial, knowing when to do it is also equally important.

The following visual depicts which algorithms need categorical data encoding and which don't.

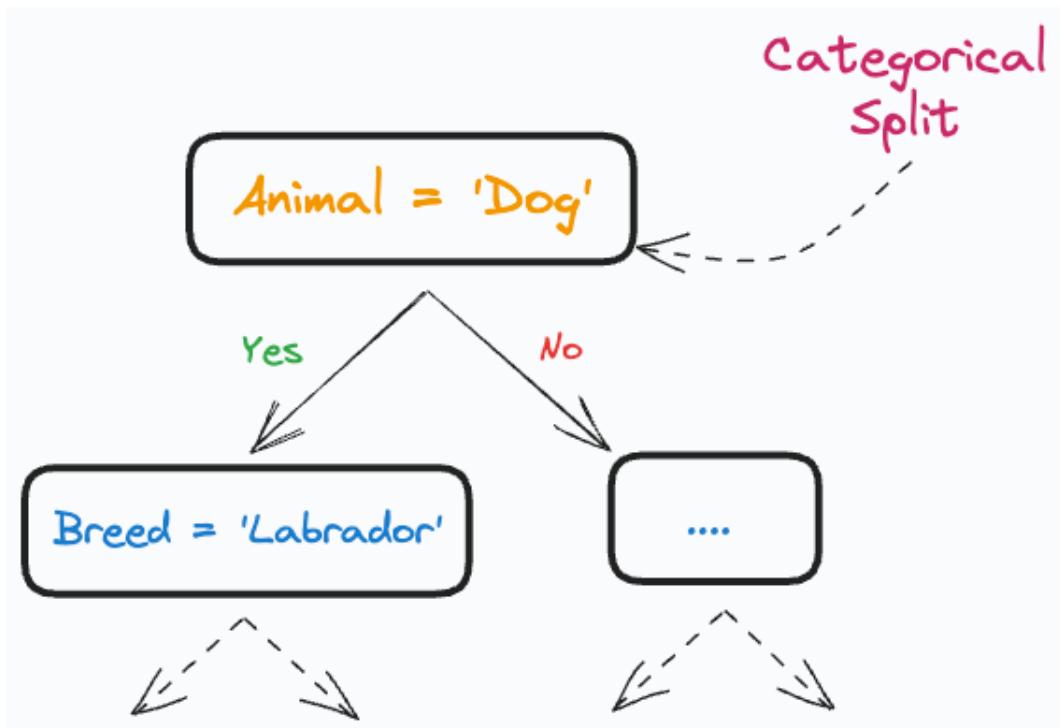




Categorization of algorithms based on categorical data encoding requirement

As shown above, many ML algorithms typically work well even without categorical data encoding. These include decision trees, random forests, naive bayes, gradient boosting, and more.

Consider a decision tree, for instance. It can split the data based on exact categorical feature values. This makes categorical feature encoding an unnecessary step.



Thus, it's important to understand the nature of your data and the algorithm you intend to use.

You may never need to encode categorical data if the algorithm is insensitive to it.

👉 Over to you: Where would you place k-nearest neighbors in this chart? Let me know :)



Scikit-LLM: Integrate Sklearn API with Large Language Models



```
# 1) Imports
from skllm.config import SKLLMConfig
from skllm import ZeroShotGPTClassifier

# 2) Set OpenAI API Key and Organisation
SKLLMConfig.set_openai_key("<YOUR_KEY>")
SKLLMConfig.set_openai_org("<YOUR_ORG>")
```

Set
Config

Use
Sklearn-like
API on LLMs

```
# 3) Define your classification dataset
X = ["Good product", "Does not work"]
y = ["positive", "negative"]

# 4) Define GPT Classifier
clf = ZeroShotGPTClassifier("gpt-3.5-turbo")

# 5) Fit Classifier
clf.fit(X, y)

# 6) Use Classifier for predictions
>>> clf.predict(X)
["positive", "negative"]
```

 DailyDoseofDS.com

Scikit-LLM is an open-source tool that offers a sklearn-compatible wrapper around OpenAI's API.

In simple words, it combines the power of LLMs with the elegance of sklearn API.

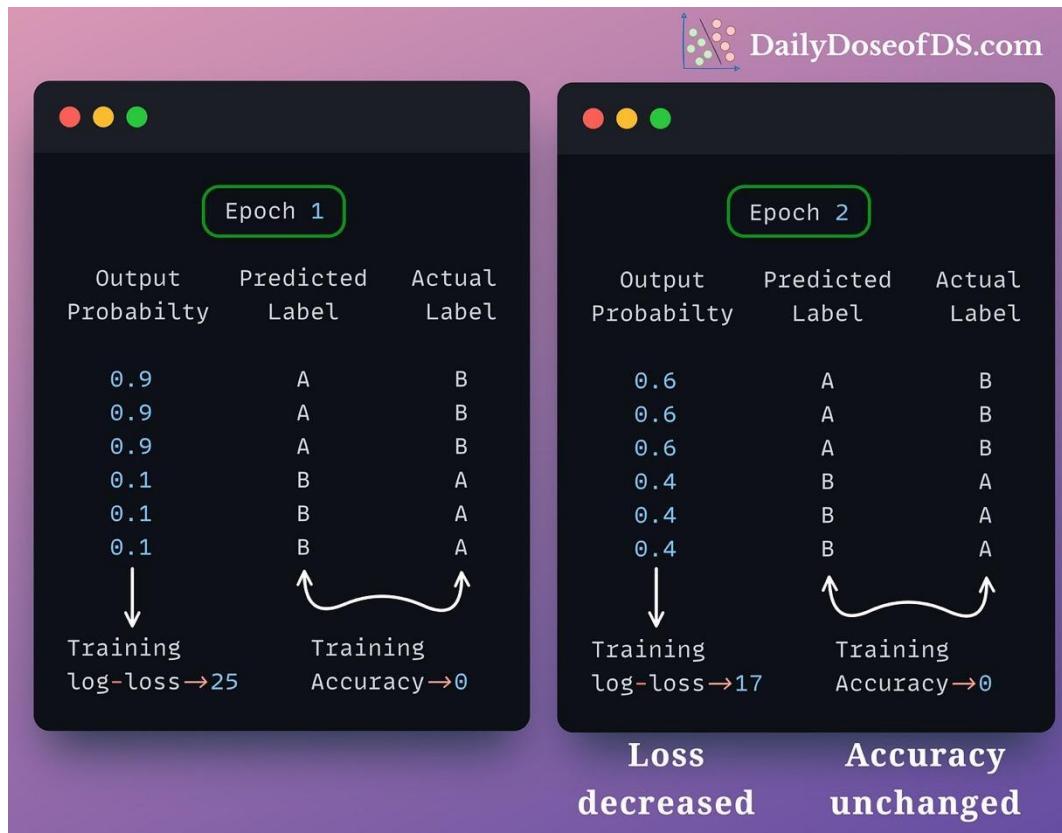
Thus, you can leverage LLMs using common sklearn functions such as fit, predict, score, etc.

What's more, you can also place LLMs in the sklearn pipeline.

Get started: [Scikit-LLM GitHub](#).



The Counterintuitive Behaviour of Training Accuracy and Training Loss



Intuitively, the training accuracy and loss are expected to be always inversely correlated.

It is expected that better predictions should lead to a lower training loss.

But that may not always be true.

In other words, you may see situations where the training loss decreases. Yet, the training accuracy remains unchanged (or even decreases).



Training loss and training accuracy decreasing



But how could that happen?

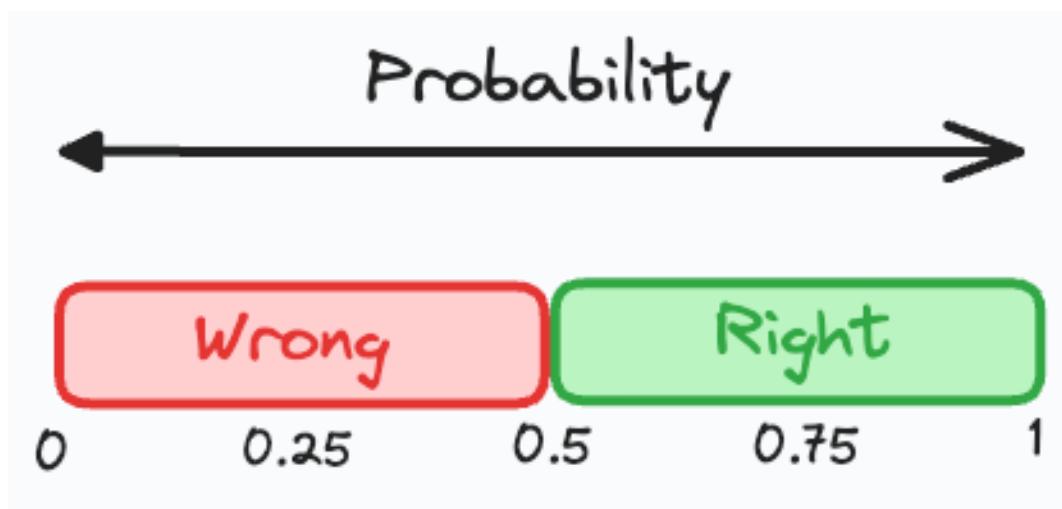
Firstly, understand that during training, the network ONLY cares about optimizing the loss function.

It does not care about the accuracy at all.

Accuracy is an external measure we introduce to measure the model's performance.

Now, when estimating the accuracy, the only thing we consider is whether we got a sample right or not.

Think of accuracy on a specific sample as a discrete measure of performance. Either right or wrong. That's it.



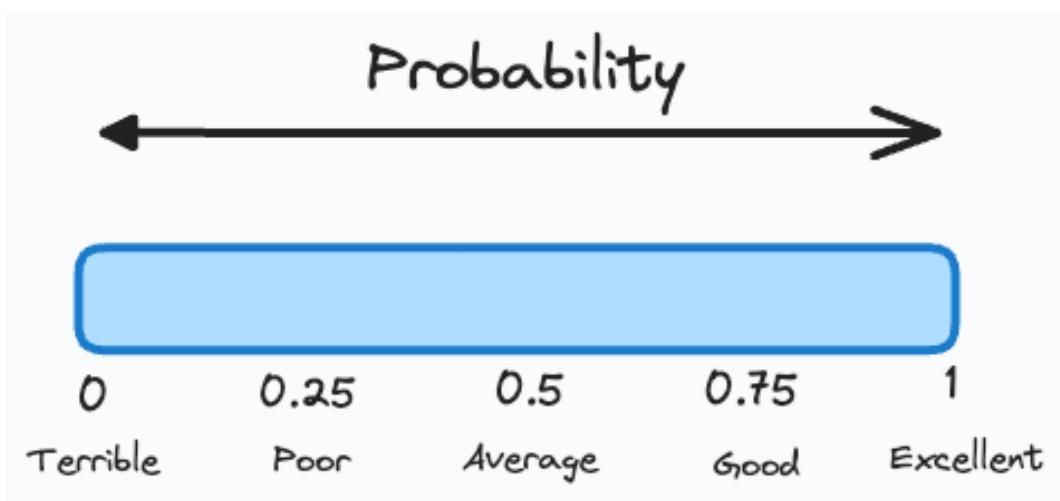
Accuracy computation

In other words, accuracy does not care if the network predicts a dog with 0.6 probability or 0.99 probability. They both have the same meaning (assuming the classification threshold is, say, 0.5).

However, the loss function is different.

It is a continuous measure of performance.

It considers how confident the model is about a prediction.



Loss spectrum

Thus, the loss function cares if the network predicted a dog with 0.6 probability or 0.99 probability.

This, at times, leads to the counterintuitive behavior of decreasing loss yet stable (or decreasing) accuracy.

If the training loss and accuracy both are decreasing, it may mean that the model is becoming:



Training loss and accuracy decreasing

More and more confident on correct predictions, and at the same time...

Less confident on its incorrect predictions.

But overall, it is making more mistakes than before.

If the training loss is decreasing, but the accuracy is stable, it may mean that the model is becoming:



Loss ↓ **Accuracy** —

Training loss decreasing, accuracy stable

More confident with its predictions. Given more time, it should improve.

But currently, it is not entirely confident to push them on either side of the probability threshold.

Having said that, remember that these kinds of fluctuations are quite normal, and you are likely to experience them.

The objective is to make you understand that while training accuracy and loss do appear to be seemingly negatively correlated, you may come across such counterintuitive situations.

Over to you: What do you think could be some other possible reasons for this behavior?



A Highly Overlooked Point In The Implementation of Sigmoid Function

```
def sigmoid(x):
    z = np.exp(-x)
    return 1/(1+z)

>>> sigmoid(+1000)    ✓
>>> sigmoid(-1000)    ✗
```

Overflow for large negative values

Use two variations of sigmoid

```
def sigmoid(x):

    if x > 0:
        z = np.exp(-x)
        return 1/(1+z)

    else:
        z = np.exp(x)
        return z/(1+z)
```

{

```
>>> sigmoid(+1000)    ✓
>>> sigmoid(-1000)    ✓
```

Prevent Overflow

There are two variations of the sigmoid function:

Standard: with an exponential term e^{-x} in the denominator only.



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Rearranged: with an exponential term e^x in both numerator and denominator.

$$\sigma(x) = \frac{e^x}{1 + e^x}$$

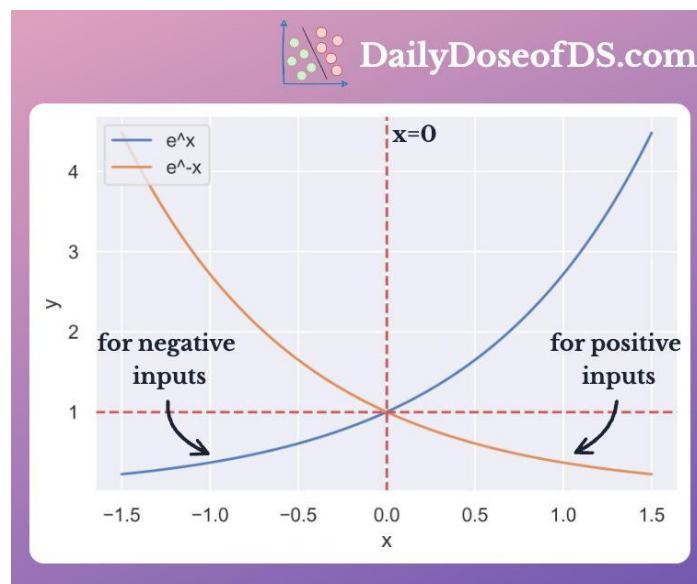
The standard sigmoid function can be easily computed for positive values. However, for large negative values, it raises overflow errors.

This is because, for large negative inputs, $e^{(-x)}$ gets bigger and bigger.

To avoid this, use both variations of sigmoid.

Standard variation for positive inputs. This prevents overflow that may occur for negative inputs.

Rearranged variation for negative inputs. This prevents overflow that may occur for positive inputs.





This way, you can maintain numerical stability by preventing overflow errors in your ML pipeline.

Having said that, luckily, if you are using an existing framework, like Pytorch, you don't need to worry about this.

These implementations offer numerical stability by default. However, if you have a custom implementation, do give it a thought.

Over to you:

- 1.** Sigmoids's two-variation implementation that I have shared above isn't vectorized. What is your solution to vectorize this?
- 2.** What are some other ways in which numerical instability may arise in an ML pipeline? How to handle them?



The Ultimate Categorization of Clustering Algorithms



Type of Clustering Algorithm	Visual Overview	Description	Algorithm(s)
Centroid-based		Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
Connectivity-based		Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
Density-based		Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
Graph-based		Cluster points based on graph distance	Affinity Propagation Spectral Clustering
Distribution-based		Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models
Compression-based		Transform data to a lower dimensional space and then perform clustering	BIRCH

Clustering is one of the core branches of unsupervised learning in ML.

It involves grouping data points together based on their inherent patterns or characteristics.

By identifying similarities (and dissimilarities) within a dataset, clustering helps in revealing underlying structures, discovering hidden patterns, and gaining insights into the data.

While centroid-based is the most common class of clustering, there's a whole world of algorithms beyond that, which we all should be aware of.

- **Centroid-based:** Cluster data points based on proximity to centroids.



- **Connectivity-based:** Cluster points based on proximity between clusters.
- **Density-based:** Cluster points based on their density.
- **Graph-based:** Cluster points based on graph distance.
- **Distribution-based:** Cluster points based on their likelihood of belonging to the same distribution.
- **Compression-based:** Transform data to a lower dimensional space and then perform clustering



Improve Python Run-time Without Changing A Single Line of Code

The diagram illustrates the performance gap between Python and Pypy. It shows a code editor with the following Python script:

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

When run in a standard Python environment, the command \$ python big_loop.py results in a run-time of 10.9s.

However, when run in a Pypy environment, the command \$ pypy big_loop.py results in a run-time of 0.41s, demonstrating a significant performance improvement of over 25x.

Python's default interpreter — CPython, isn't smart at dealing with for-loops, lists, and more.

It serves as a standard interpreter for Python and offers no built-in optimization.

Pypy, however, is an alternative implementation of CPython, which is much smarter.

How does it work?

It takes existing Python code and generates fast machine code using just-in-time compilation.

As a result, post compilation, the code runs at native machine code speed.



And Pypy can be used without modifying a single line of existing Python code.

👉 You should consider Pypy when:

- you're dealing with standard Python objects.
- speedups aren't possible with NumPy/Pandas.

So remember...

When you have some native Python code, don't run it with the default interpreter of Python.

Instead, look for alternative smarter implementations, such as Pypy.

Pypy will help you:

- improve run-time of code, and
- execute it at machine speed,
- without modifying the code.

Find some more benchmarking results between Python and Pypy below:

The screenshot shows two side-by-side terminal windows on a Mac OS X desktop. Both windows have a dark theme and a purple gradient background.

Top Left Terminal (Python):

```
def fib(N):
    """
    Function to find the
    Nth Fibonacci number.

    fib(N) = fib(N-1) + fib(N-2)
    ...

```

Top Right Terminal (Pypy):

```
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 (Python):

```
$ python fib.py # N=35
# Time: 2.53s
```

```
$ python fib.py # N=45
# Time: 296s
```

```
$ python pi.py # n_terms=10^8
# Time: 14.7s
```

Bottom Right Terminal (Pypy):

```
$ pypy fib.py # N=35
# Time: 0.08s (~30x Faster)
```

```
$ pypy fib.py # N=45
# Time: 11.2s (~27x Faster)
```

```
$ pypy pi.py # n_terms=10^8
# Time: 0.49s (~30x Faster)
```

Arrows point from the bottom-left terminal's output lines to the corresponding faster execution times in the bottom-right terminal's output. The arrows are white with black outlines.

Get started with Pypy here: [Pypy docs](#).



A Lesser-Known Feature of the Merge Method in Pandas

avichawla.substack.com

```
df1 = pd.DataFrame({'colA': ['A', 'B', 'C'],
                     'colB': [1, 2, 3]})

df2 = pd.DataFrame({'colA': ['A', 'B', 'C'],
                     'colC': [4, 5, 6]})

pd.merge(df1, df2, on = 'colA', validate='one_to_one')
```

Unique keys

	colA	colB	colC
0	A	1	4
1	B	2	5
2	C	3	6

One-one Merge Validated! No Merge Error

```
df1 = pd.DataFrame({'colA': ['A', 'B', 'C'],
                     'colB': [1, 2, 3]})

df2 = pd.DataFrame({'colA': ['A', 'B', 'B'],
                     'colC': [4, 5, 6]}) ← Repeated keys

pd.merge(df1, df2, on = 'colA', validate='one_to_one')
```

MergeError: Merge keys are not unique in right dataset; not a one-to-one merge
One-one Merge Invalidated!

When merging two DataFrames, one may want to perform some checks to ensure the integrity of the merge operation.

For instance, if one of the two DataFrames has repeated keys, it will result in duplicated rows.

But this may not be desired.

The good thing is that you can check this with Pandas.

Pandas' `merge()` method provides a `validate` parameter, which checks if the merge is of a specified type or not.

- “one_to_one”: Merge keys should be unique in both DataFrames.
- “one_to_many”: Merge keys should be unique in the left DataFrame.
- “many_to_one”: Merge keys should be unique in the right DataFrame.
- “many_to_many”: Merge keys may or may not be unique in both DataFrames.



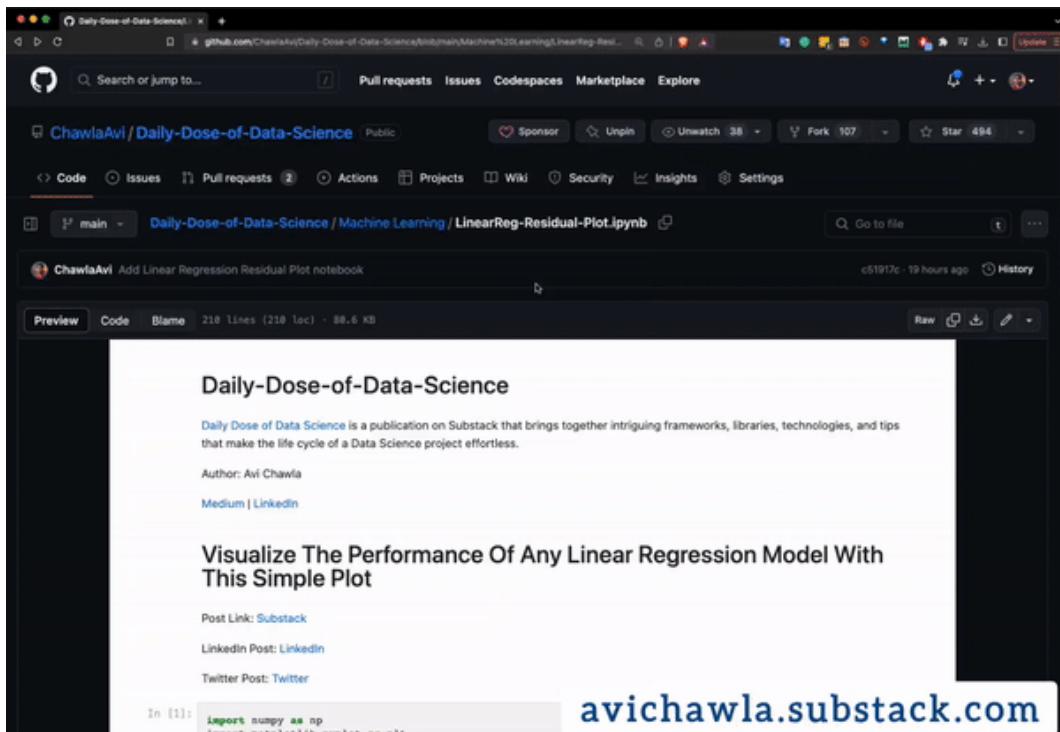
Pandas raises a **Merge Error** if the merge operation does not conform to the specified type.

This helps you prevent errors that can often go unnoticed.

Over to you: What are some other hidden treasures you know of in Pandas? Let me know :)



The Coolest GitHub-Colab Integration You Would Ever See



If you have opened a notebook in GitHub, change the domain from "**github**" to "**githubbtocolab**".

As a result, the notebook will directly open in Google Colab. Isn't that cool?

Try it out today.

Open any Jupyter Notebook hosted on GitHub, and change the domain as suggested above.



Most Sklearn Users Don't Know This About Its LinearRegression Implementation

```
>>> X.shape  
(100000, 2000) # Large no. of features  
  
# Train LinearReg and SGDReg  
>>> lin_reg = LinearRegression().fit(X, y)  
>>> sgd_reg = SGDRegressor(...).fit(X, y)  
  
# Model Weights (first 5 features):  
lin_reg → [-0.36, -0.45, -0.62, 0.76, 0.84]  
sgd_reg → [-0.37, -0.45, -0.62, 0.77, 0.85]  
  
# Model Intercept:  
lin_reg → 0.00241  
sgd_reg → 0.00237  
  
# Training Run-time:  
lin_reg → 51 seconds  
★ sgd_reg → 18 seconds ~3x Faster
```

Similar weights

Similar intercept

★ ~3x Faster

Sklearn's `LinearRegression` class implements the ordinary least square (OLS) method to find the best fit.

Some important characteristics of OLS are:

- It is a deterministic algorithm. If run multiple times, it will always converge to the same weights.
- It has no hyperparameters.
- It involves matrix inversion, which is cubic in relation to the no. of features. This gets computationally expensive with many features.

Read this answer to learn more about OLS' run-time complexity: [StackOverflow](#).



SGDRegressor, however:

- is a stochastic algorithm. It finds an approximate solution using optimization.
- has hyperparameters.
- involves gradient descent, which is relatively inexpensive.

Now, if you have many features, Sklearn's LinearRegression will be computationally expensive.

This is because it relies on OLS, which involves matrix inversion. And as mentioned above, inverting a matrix is cubic in relation to the total features.

This explains the run-time improvement provided by Sklearn's SGDRegressor over LinearRegression.

So remember...

When you have many features, avoid using Sklearn's LinearRegression.

Instead, use the [SGDRegressor](#).

- This will help you:
- Improve run-time.
- Avoid memory errors.

Implement batching (if needed). I have covered this in one of my previous posts here: [A Lesser-Known Feature of Sklearn To Train Models on Large Datasets](#).

👉 Over to you: What are some tradeoffs between using LinearRegression vs. SGDRegressor?



Break the Linear Presentation of Notebooks With Stickyland

The screenshot illustrates the Stickyland extension for JupyterLab. It features four main components:

- A Notebook Cell:** A standard Jupyter notebook cell containing code and output.
- B Sticky Dock:** A dashboard area where a cell from the notebook can be moved to. It includes a "New" button, a placeholder for dragging cells, and a "Create" button.
- C Sticky Cell:** A floating window containing the same code and output as the original cell, with its own toolbar.
- D Floating Cell:** Another floating window showing the full dataset, with an "auto-run" toggle switch.

The background of the interface is purple, and the overall theme is "STICKYLAND".

If you are a JupyterLab user, here's a pretty interesting extension for you.

Stickyland is an open-source tool that lets you break the linear presentation of a notebook.

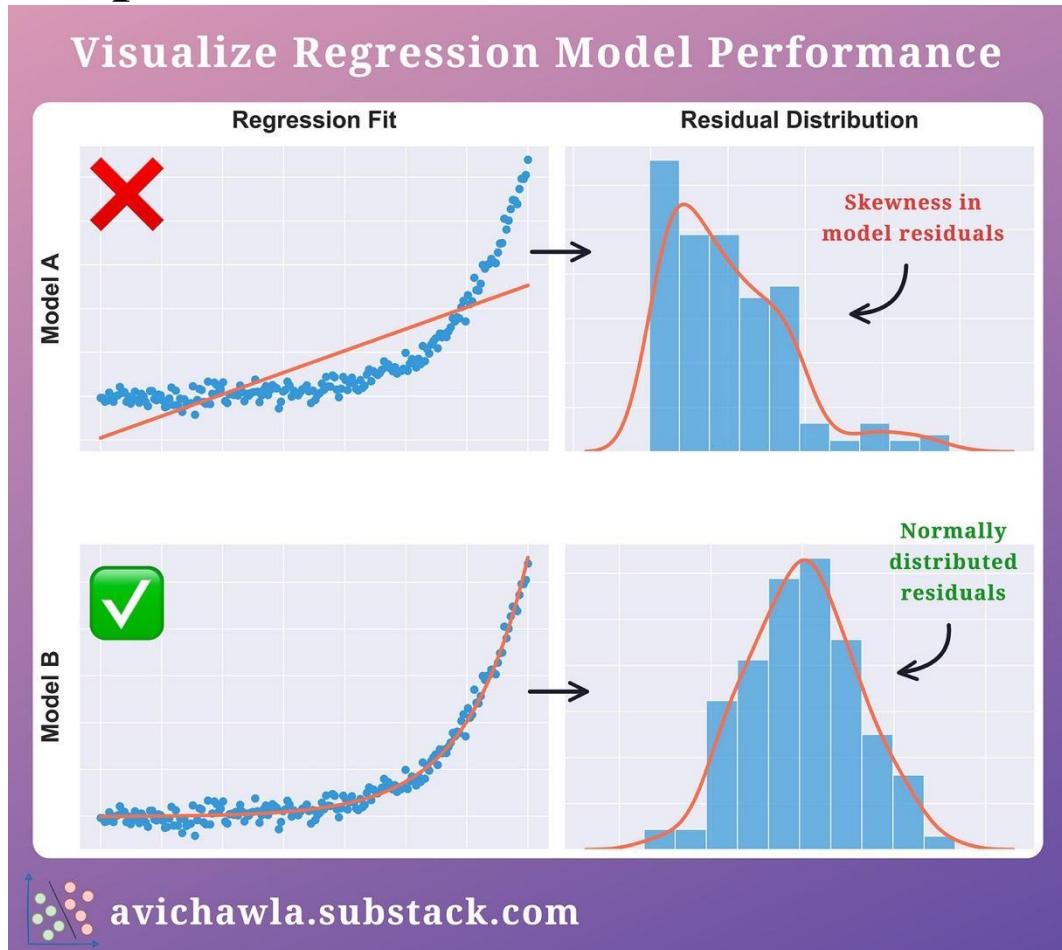
With Stickyland, you can:

- Create floating cells (code and markdown)
- Auto-run cells when any changes take place
- Take notes in Jupyter
- Organize the notebook as a dashboard

Find more info about Stickyland here: [GitHub](#) | [Paper](#).



Visualize The Performance Of Any Linear Regression Model With This Simple Plot



Linear regression assumes that the model residuals (=actual-predicted) are normally distributed.

If the model is underperforming, it may be due to a violation of this assumption.

A residual distribution plot is a great way to verify this and also determine the model's performance.

As the name suggests, it depicts the distribution of residuals (=actual-predicted).

A good residual plot will:

- Follow a normal distribution
- NOT reveal trends in residuals

A bad residual plot will:



- Show skewness
- Reveal patterns in residuals

Thus, the more normally distributed the residual plot looks, the more confident you can be about your model.

This is especially useful when the regression line is difficult to visualize, i.e., in a high-dimensional dataset.

So remember...

After running a linear model, always check the distribution of the residuals.

This will help you:

- Validate the model's assumptions
- Determine how good your model is
- Find ways to improve it (if needed)

👉 Over to you: What are some other ways/plots to determine the linear model's performance?

Thanks for reading!



Waterfall Charts: A Better Alternative to Line/Bar Plot



To visualize the change in value over time, 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, at times, it can be difficult to visually estimate the scale of incremental changes.

Instead, create a waterfall chart. It elegantly depicts these rolling differences.

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.

Over to you: What are some other cases where Bar/Line plots aren't an ideal choice?

Read more here: [GitHub](#).



What Does The Google Styling Guide Say About Imports

The diagram illustrates a comparison between two ways of importing functions from a module. On the left, a red box highlights a snippet where a function is imported directly:

```
from my_package.my_module  
import my_function  
  
>>> my_function() → ✗
```

A white arrow points from this snippet to the text "Don't import function" on the right.

On the right, a green box highlights a snippet where the entire module is imported:

```
from my_package import my_module  
  
>>> my_module.my_function() → ✓
```

A white arrow points from the text "Import module instead" on the left towards this snippet.

Recently, I was reviewing Google's Python style guide. Here's what it says about imports.

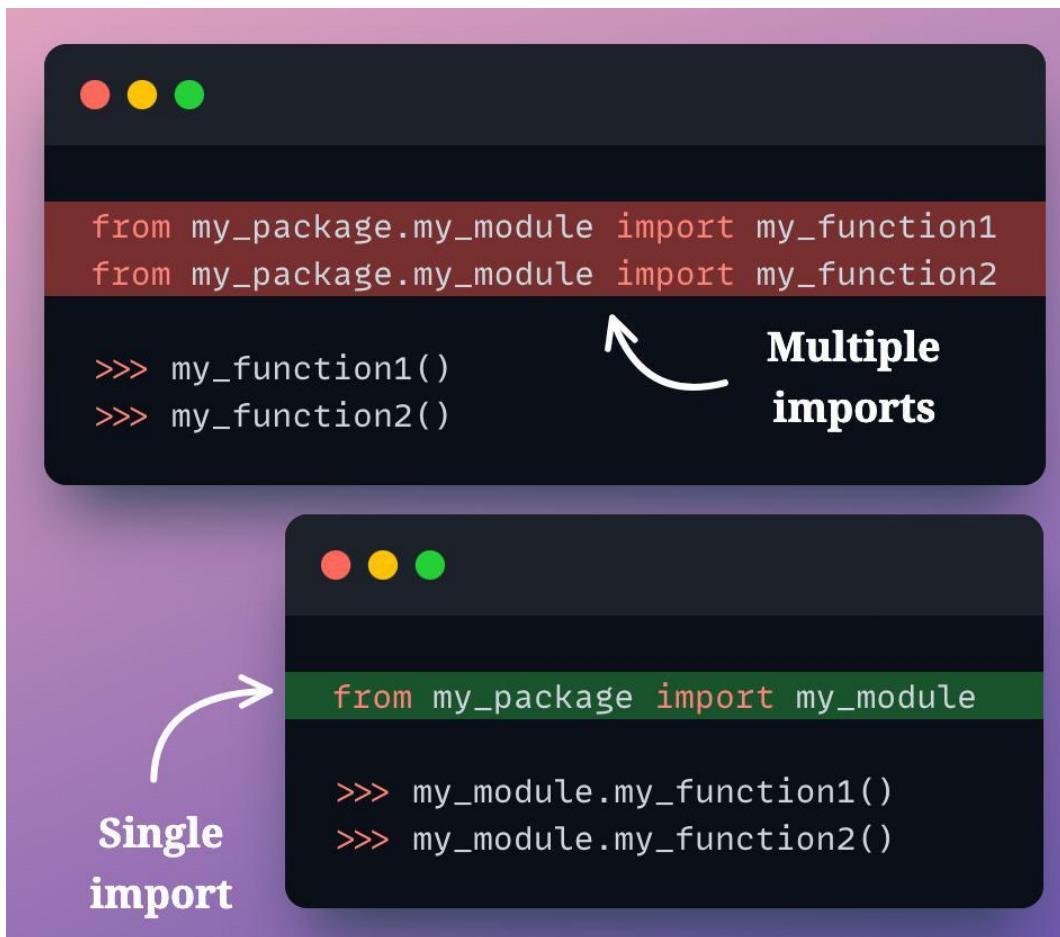
When writing import statements, it is recommended to import the module instead of a function.

Quoting from the style guide:

Use import statements for packages and modules only, not for individual classes or functions.

Advantages:

- 1. Namespace clarity:** Importing the module makes the source of the function clear during invocation. For instance, `my_module.my_function()` tells that `my_function` is defined in `my_module`.
- 2. Fewer import statements:** If you intend to use many functions from a specific module, importing each function may not be feasible, as shown below:



The diagram illustrates two code snippets side-by-side. The top snippet shows 'Multiple imports' where a module is imported twice:

```
from my_package.my_module import my_function1
from my_package.my_module import my_function2

>>> my_function1()
>>> my_function2()
```

The bottom snippet shows a 'Single import' where the entire module is imported once, and functions are called using the module name prefix:

```
from my_package import my_module

>>> my_module.my_function1()
>>> my_module.my_function2()
```

Curved arrows point from the text labels 'Multiple imports' and 'Single import' to their respective code examples.

Disadvantage:

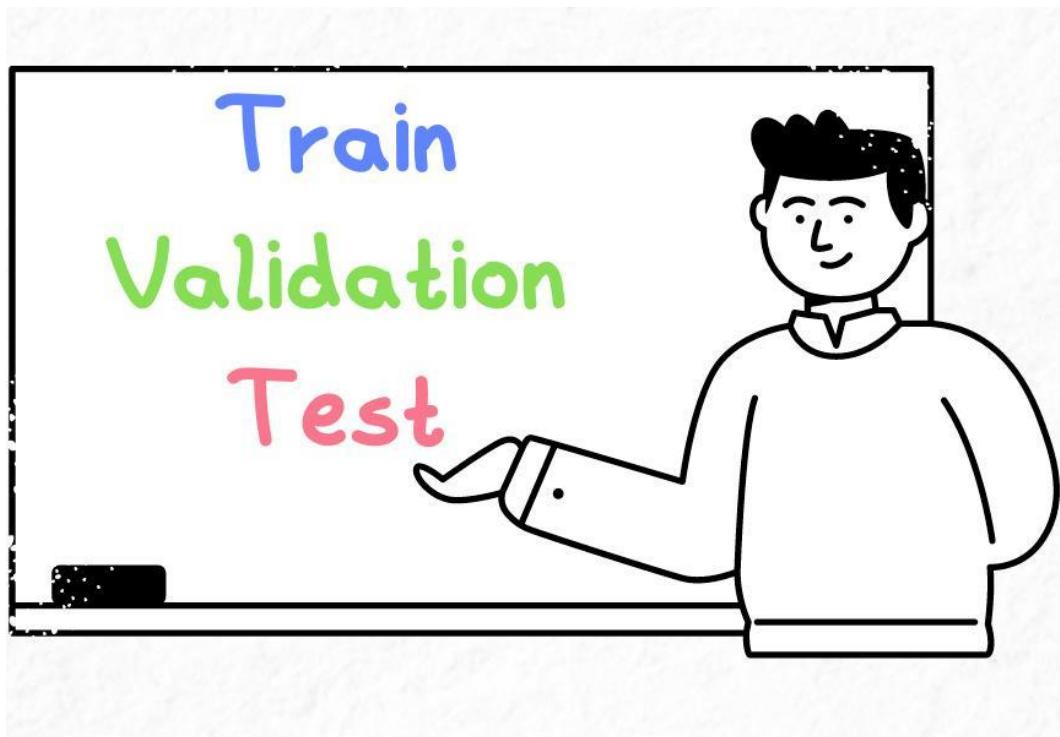
- **Explicitness:** When you import a function instead of a module, it's immediately clear which functions you intend to use from the module.

Over to you: While the guideline is intended for Google's internal stuff, it may be applicable elsewhere. Yet, when would you still prefer importing functions over modules? Let me know :)

Read the full guide here: [Google style guide](#).



How To Truly Use The Train, Validation and Test Set



Everyone knows about the train, test, and validation sets. But very few understand how to use them correctly.

Here's what you should know about splitting data and using it for ML models.

Begin by splitting the data into:

- Train
- Validation
- Test

Now, assume that the test data does not even exist. Forget about it instantly.

Begin with the train set. This is your whole world now.

- Analyze it
- Transform it
- Model it

As you finish modeling, you'd want to measure the model's performance on unseen data.

Bring in the validation set now.



Based on validation performance, improve the model.

Here's how you iteratively build your model:

- Train using a train set
- Evaluate it using the validation set
- Improve the model
- Evaluate again using the validation set
- Improve the model again
- and so on.

Until...

You start overfitting the validation set. This indicates that you exploited and polluted the validation set.

No worries.

Merge it with the train set and generate a new split of train and validation.

Note: Rely on cross-validation if needed, especially when you don't have much data. You may still use cross-validation if you have enough data. But it can be computationally intensive.

Now, if you are happy with the model's performance, evaluate it on test data.

👉 What you use a test set for:

Get a final and unbiased review of the model.

👉 What you DON'T use a test set for:

Analysis, decision-making, etc.

If the model is underperforming on the test set, no problem. Go back to the modeling stage and improve it.

BUT (and here's what most people do wrong)!

They use the same test set again.

Not allowed!

Think of it this way.

Your professor taught you in class. All in-class lessons and examples are the train set.

The professor gave you take-home assignments, which act like validation sets.



You get some wrong and some right.

Based on this, you adjust your topic fundamentals, i.e., improve the model.

If you keep solving the same take-home assignment, then you will eventually overfit it, won't you?

That is why we bring in a new validation set after some iterations.

The final exam day paper is your test set.

If you do well, awesome!

But if you fail, the professor cannot give you the same exam paper next time, can they? You know what's inside.

That is why we always use a specific test set only ONCE.

Once you do, merge it with the train and validation set and generate a new split.

Repeat.



Restart Jupyter Kernel Without Losing Variables

The screenshot shows a Jupyter Notebook interface with the following code cells:

```
In [ ]: # Define variable  
value = 2
```

```
In [ ]: # Store it using magic command  
%store value
```

Restart Kernel

```
In [ ]: # Restore variable  
%store -r value
```

```
In [ ]: value
```

```
In [ ]:
```

avichawla.substack.com

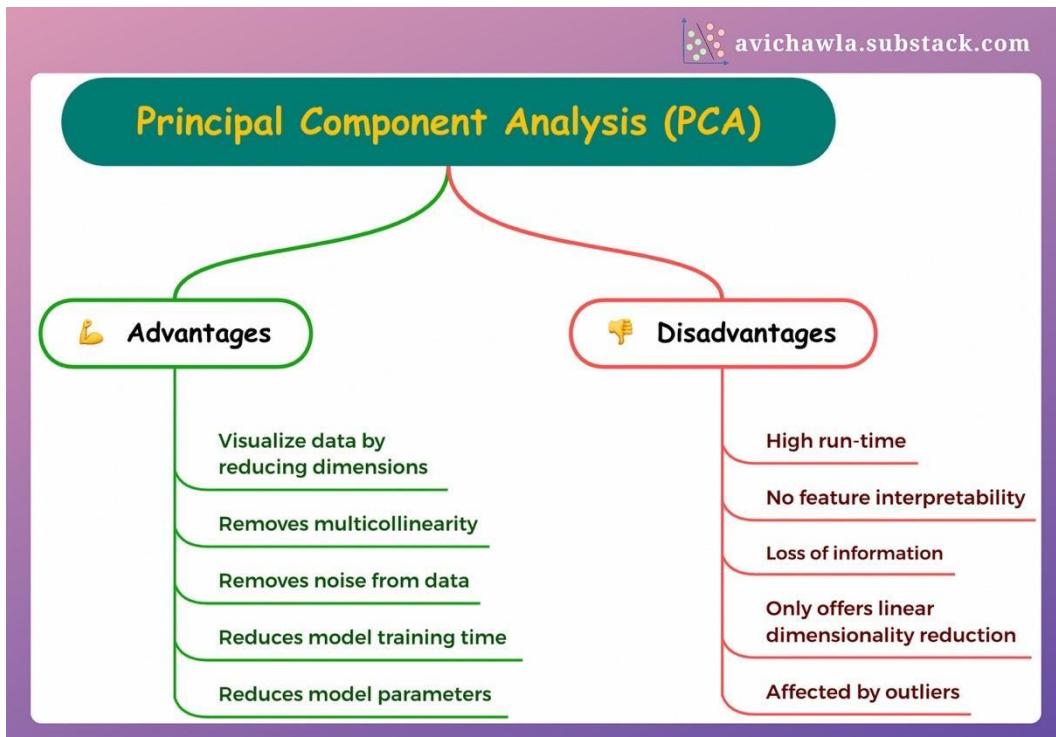
While working in a Jupyter Notebook, you may want to restart the kernel due to several reasons. Instead of dumping the variables to disk before restarting Jupyter and reading them back, use the **store magic command**.

It allows you to store and retrieve a variable back even if you restart the kernel. This avoids the hassle of dumping the object to disk.

Over to you: What are some other cool Jupyter hacks that you know of?



The Advantages and Disadvantages of PCA To Consider Before Using It



PCA is possibly the most popular dimensionality reduction technique.

If you wish to know how PCA works, I have a highly simplified post here: [A Visual and Overly Simplified Guide to PCA](#).

Yet, it is equally important to be aware of what we get vs. what we compromise when we use PCA.

The above visual depicts five common pros and cons of using PCA.

Advantages

By reducing the data to two dimensions, you can easily visualize it.

PCA removes multicollinearity. Multicollinearity arises when two features are correlated. PCA produces a set of new orthogonal axes to represent the data, which, as the name suggests, are uncorrelated.

PCA removes noise. By reducing the number of dimensions in the data, PCA can help remove noisy and irrelevant features.

PCA reduces model parameters: PCA can help reduce the number of parameters in machine learning models.



PCA reduces model training time. By reducing the number of dimensions, PCA simplifies the calculations involved in a model, leading to faster training times.

Disadvantages

The run-time of PCA is cubic in relation to the number of dimensions of the data. This can be computationally expensive at times for large datasets.

$$\text{Runtime} : O(nd^2 + d^3)$$

d : dimensions

n : samples

PCA transforms the original input variables into new principal components (or dimensions). The new dimensions offer no interpretability.

While PCA simplifies the data and removes noise, it always leads to some loss of information when we reduce dimensions.

PCA is a linear dimensionality reduction technique, but not all real-world datasets may be linear. Read more about this in my previous post here: [The Limitation of PCA Which Many Folks Often Ignore](#).

PCA gets affected by outliers. This can distort the principal components and affect the accuracy of the results.



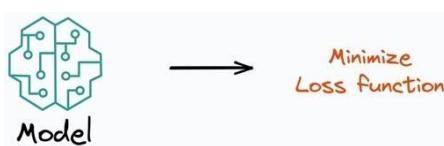
Loss Functions: An Algorithm-wise Comprehensive Summary



Algorithm	Commonly Used Loss Function or Training Methodology
Linear Regression	Mean Squared Error
Logistic Regression	Cross-Entropy Loss
Decision Tree Classifier	Information Gain or Gini impurity
Decision Tree Regressor	Mean Squared Error
Random Forest Classifier	Information Gain or Gini impurity
Random Forest Regressor	Mean Squared Error
Support Vector Machines (SVMs)	Hinge Loss
k-Nearest Neighbors	No loss function as kNN is non-parameteric
Naive Bayes	No loss function
Neural Networks	Regression: Mean Squared Error Classification: Cross-Entropy Loss
AdaBoost	Exponential loss
Gradient Boosting LightGBM CatBoost XGBoost	Regression: Mean Squared Error Classification: Cross-Entropy Loss
KMeans Clustering	No loss function as KMeans is unsupervised

Loss functions are a key component of ML algorithms.

They specify the objective an algorithm should aim to optimize during its training. In other words, loss functions tell the algorithm what it should be trying to minimize or maximize in order to improve its performance.





Therefore, knowing which loss functions are best suited for specific ML algorithms is extremely crucial.

The above visual depicts the most commonly used loss functions with various ML algorithms.

1. Linear Regression: Mean Squared Error (MSE). This can be used with and without regularization, depending on the situation.

2. Logistic regression: Cross-Entropy Loss or Log Loss, with and without regularization.

3. Decision Tree and Random Forest:

- a. Classifier: Gini impurity or information gain.
- b. Regressor: Mean Squared Error (MSE)

4. Support Vector Machines (SVMs): Hinge loss. Read more: [Wikipedia](#).

5. k-Nearest Neighbors (kNN): No loss function. kNN is a non-parametric lazy learning algorithm. It works by retrieving instances from the training data, and making predictions based on the k nearest neighbors to the test data instance.

6. Naive Bayes: No loss function. Can you guess why? Let me know in the comments if you need help.

7. Neural Networks: They can use a variety of loss functions depending on the type of problem. The most common are:

- a. Regression: Mean Squared Error (MSE).
- b. Classification: Cross-Entropy Loss.

8. AdaBoost: Exponential loss function. AdaBoost is an ensemble learning algorithm. It combines multiple weak classifiers to form a strong classifier. In each iteration of the algorithm, AdaBoost assigns weights to the misclassified instances from the previous iteration. Next, it trains a new weak classifier and minimizes the weighted exponential loss.

9. Other Boosting Algorithms:

- a. Regression: Mean Squared Error (MSE).
- b. Classification: Cross-Entropy Loss.

10. KMeans Clustering: No loss function.

Over to you: Which algorithms have I missed?



Is Data Normalization Always Necessary Before Training ML Models?

Data normalization is commonly used to improve the performance and stability of ML models.

This is because normalization scales the data to a standard range. This prevents a specific feature from having a strong influence on the model's output. What's more, it ensures that the model is more robust to variations in the data.



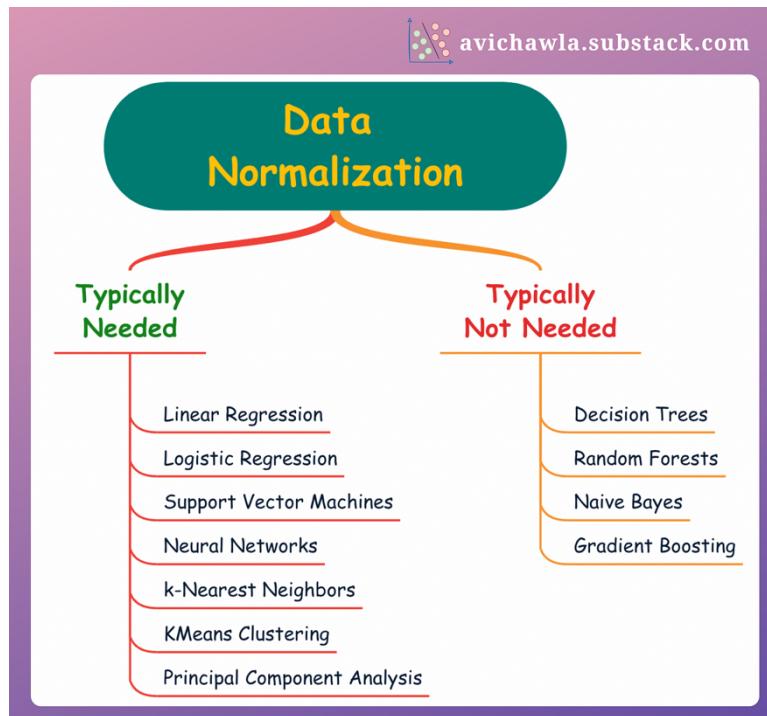
Different scales of columns

For instance, in the image above, the scale of **Income** could massively impact the overall prediction. Normalizing the data by scaling both features to the same range can mitigate this and improve the model's performance.

But is it always necessary?

While normalizing data is crucial in many cases, knowing when to do it is also equally important.

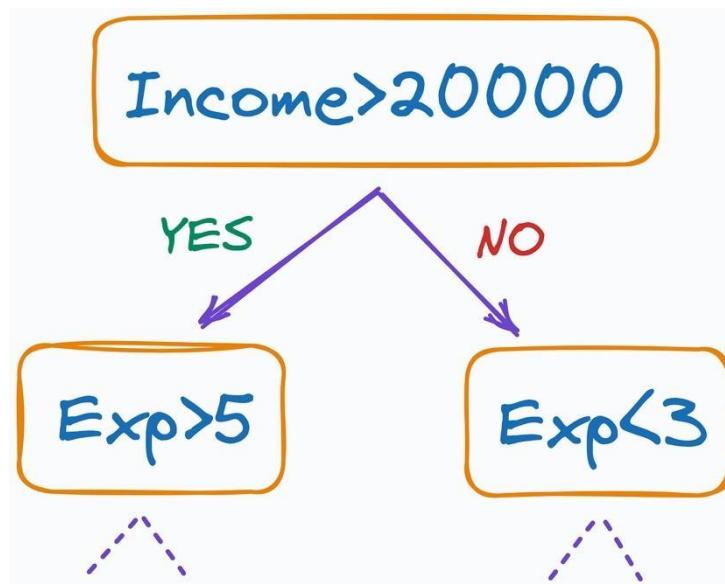
The following visual depicts which algorithms typically need normalized data and which don't.



Categorization of algorithms based on normalized data requirement

As shown above, many algorithms typically do not need normalized data. These include decision trees, random forests, naive bayes, gradient boosting, and more.

Consider a decision tree, for instance. It splits the data based on thresholds determined solely by the feature values, regardless of their scale.



Decision tree



Thus, it's important to understand the nature of your data and the algorithm you intend to use.

You may never need data normalization if the algorithm is insensitive to the scale of the data.

Over to you: What other algorithms typically work well without normalizing data?
Let me know :)



Annotate Data With The Click Of A Button Using Pigeon

The screenshot shows a Jupyter notebook cell with Python code using the `pigeon` library to annotate images. The code defines an `annotate` function with parameters: `images_list`, `options` (containing 'cat' and 'dog'), and `display_fn` (a lambda function that displays the image). Below the code, there's a row of three buttons: 'cat', 'dog', and 'skip'. Underneath the buttons is a large image of a brown tabby cat. At the bottom of the cell, there's another code block showing the annotated labels: `[(./cats and dogs/image1.jpeg, 'cat'), (./cats and dogs/image2.jpeg, 'dog'), (./cats and dogs/image3.jpeg, 'dog')]`.

```
from pigeon import annotate

annotations = annotate(images_list,
                      options=['cat', 'dog'],
                      display_fn=lambda image_file: display(Image(image_file)) ## display each example
)

cat      dog      skip

annotations ## get labels
[('./cats and dogs/image1.jpeg', 'cat'),
 ('./cats and dogs/image2.jpeg', 'dog'),
 ('./cats and dogs/image3.jpeg', 'dog')]
```

To perform supervised learning, you need labeled data. But if your data is unlabeled, you must spend some time annotating/labeling it.

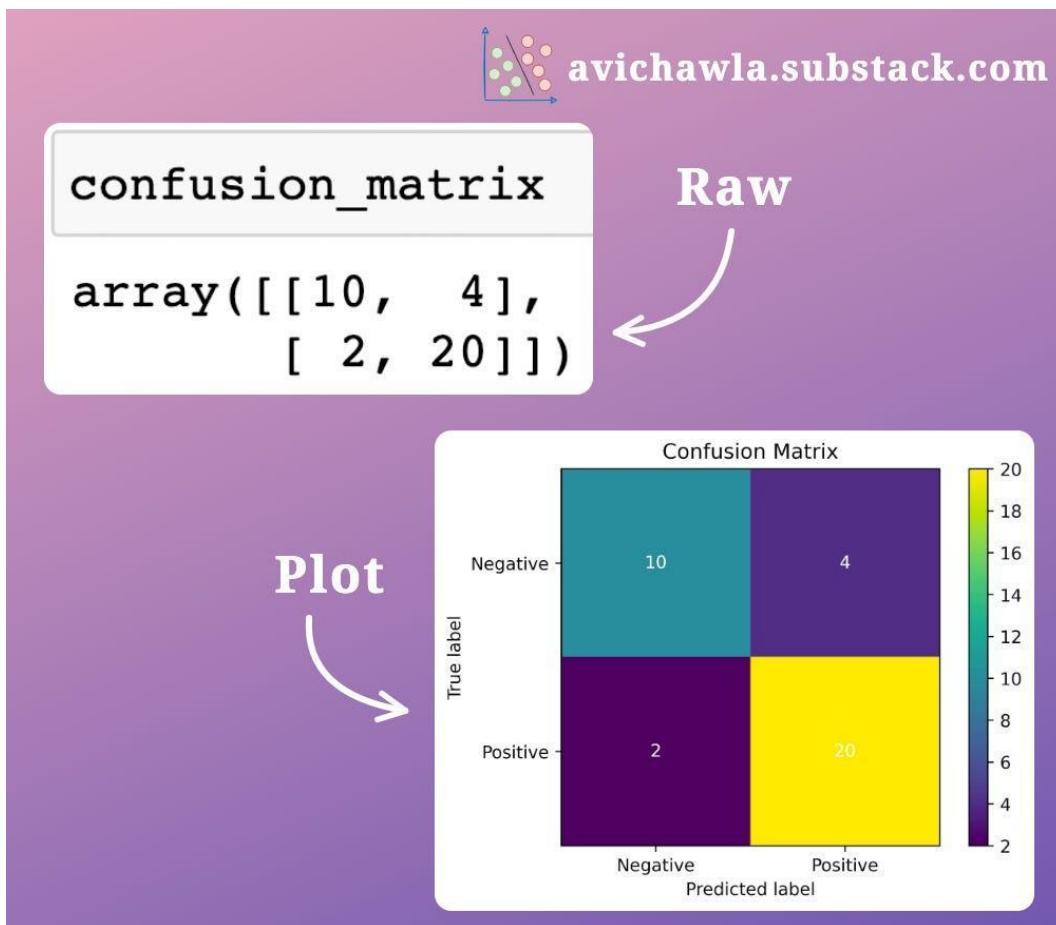
To do this quickly in a Jupyter notebook, use Pigeon. With this, you can annotate your data by simply defining buttons and labeling samples.

Read more: [Pigeon GitHub](#).



Enrich Your Confusion Matrix With A Sankey Diagram

A confusion matrix is mostly interpreted as is, i.e., raw numbers. Sometimes though, it is also visualized by plotting it.



Traditional ways of previewing confusion matrix

Yet, both these ways are not interactive and truly elegant.

Plotting a confusion matrix as a Sankey diagram is an option worth exploring here.

If you wish to read more about Sankey Diagrams, read my previous post here: [Analyse Flow Data With Sankey Diagrams](#).

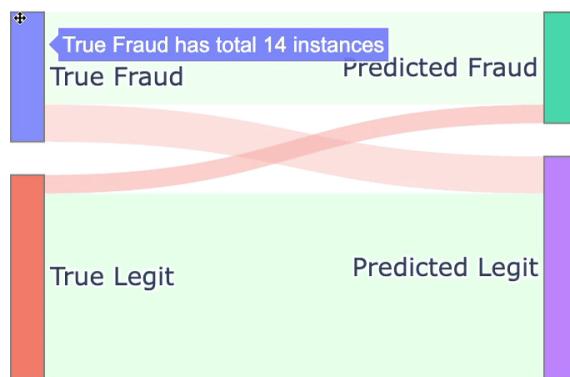
As demonstrated below, one can interactively interpret the number of instances belonging to each class and how they were classified.



```
confusion_matrix = np.array([[10, 4],  
                            [2, 20]])  
  
plot_confusion_matrix_as_sankey(confusion_matrix, ['Fraud', 'Legit'])
```



Confusion Matrix Sankey Diagram



Confusion matrix as Sankey Diagram

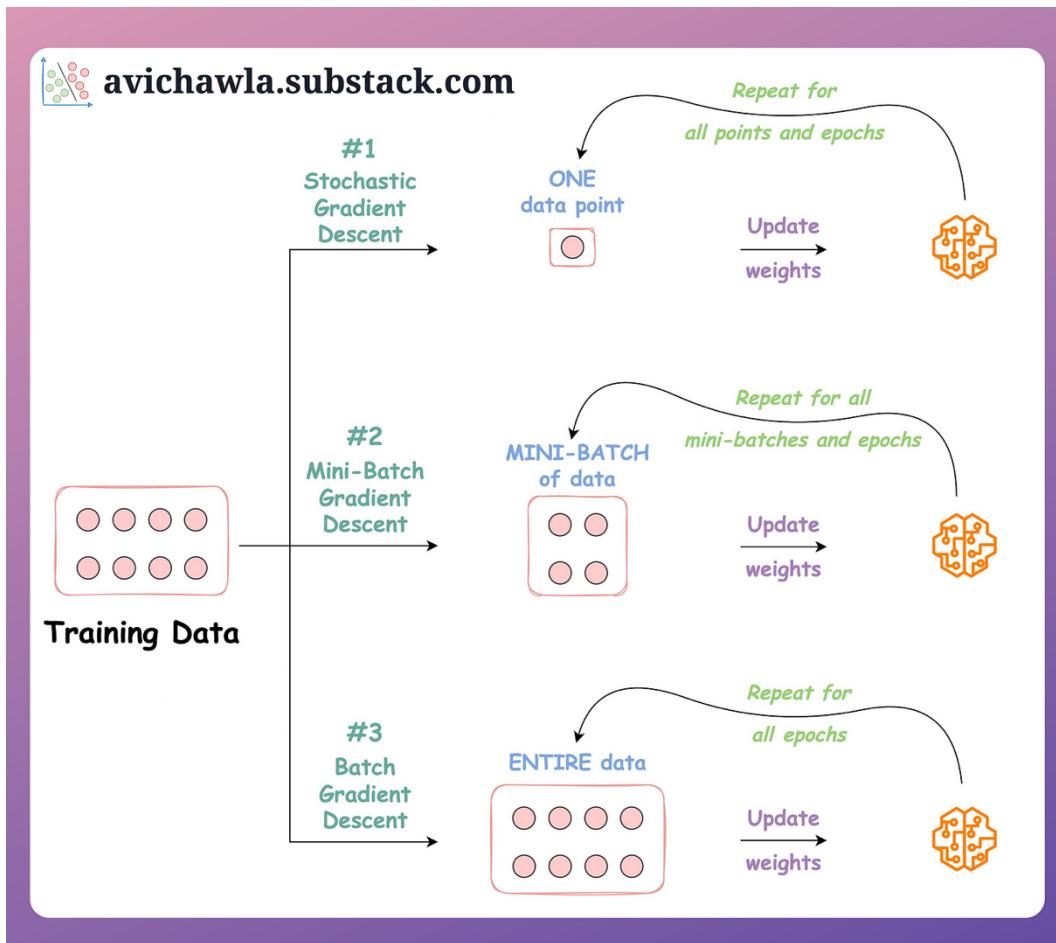
What's more, hovering over the links gives more info about those instances, which can offer better interpretability.

What do you think? Is this a better approach than the traditional ones? Let me know :)

Find the code for creating the above Confusion Matrix Sankey diagram here: [Notebook](#).



A Visual Guide to Stochastic, Mini-batch, and Batch Gradient Descent

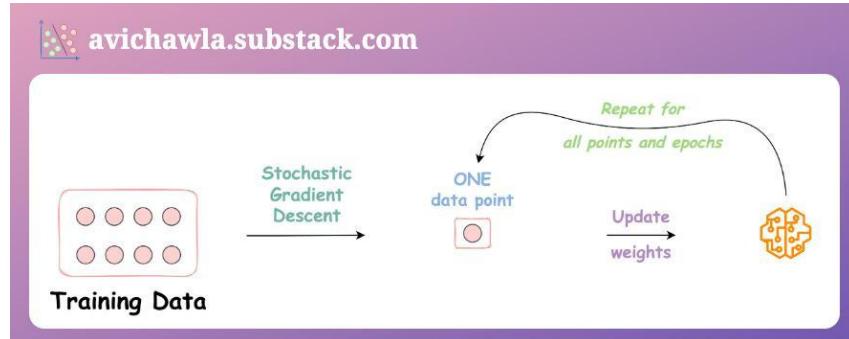


Gradient descent is a widely used optimization algorithm for training machine learning models.

Stochastic, mini-batch, and batch gradient descent are three different variations of gradient descent, and they are distinguished by the number of data points used to update the model weights at each iteration.



- ◆ Stochastic gradient descent: Update network weights using one data point at a time.



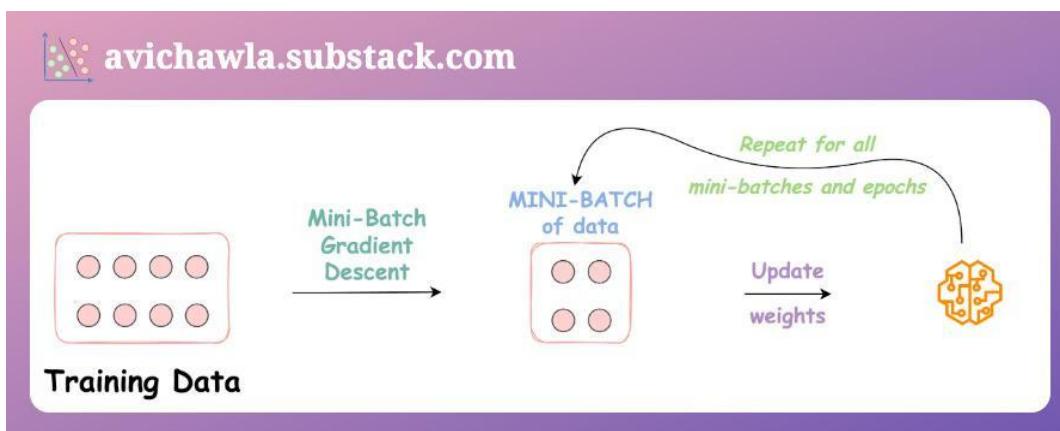
- **Advantages:**

- Easier to fit in memory.
- Can converge faster on large datasets and can help avoid local minima due to oscillations.

- **Disadvantages:**

- Noisy steps can lead to slower convergence and require more tuning of hyperparameters.
- Computationally expensive due to frequent updates.
- Loses the advantage of vectorized operations.

- ◆ Mini-batch gradient descent: Update network weights using a few data points at a time.



- **Advantages:**

- More computationally efficient than batch gradient descent due to vectorization benefits.

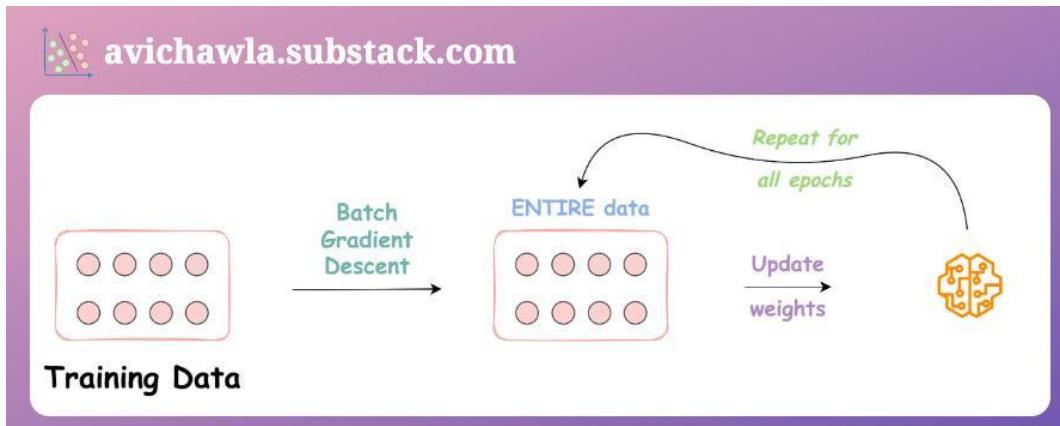


- Less noisy updates than stochastic gradient descent.

- **Disadvantages:**

- Requires tuning of batch size.
- May not converge to a global minimum if the batch size is not well-tuned.

◆ Batch gradient descent: Update network weights using the entire data at once.



- **Advantages:**

- Less noisy steps taken towards global minima.
- Can benefit from vectorization.
- Produces a more stable convergence.

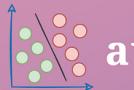
- **Disadvantages:**

- Enforces memory constraints for large datasets.
- Computationally slow as many gradients are computed, and all weights are updated at once.

Over to you: What are some other advantages/disadvantages you can think of? Let me know :)



A Lesser-Known Difference Between For-Loops and List Comprehensions



Value
changed

```
>>> a = 1  
  
>>> for a in range(6):  
    pass  
  
>>> print(a)  
5 # Output
```

Value
unchanged

```
>>> a = 1  
  
>>> [... for a in range(6)]  
  
>>> print(a)  
1 # Output
```

In the code above, the for-loop updated the existing variable (`a`), but list comprehension didn't. Can you guess why? Read more to know.

A loop variable is handled differently in for-loops and list comprehensions.

A for-loop leaks the loop variable into the surrounding scope. In other words, once the loop is over, you can still access the loop variable.

We can verify this below:



```
>>> for loop_var in range(6):
... 
>>> print(loop_var) ## Loop variable accessible
5
```

No error

In the main snippet above, as the loop variable (`a`) already existed, it was overwritten in each iteration.

But a list comprehension does not work this way. Instead, the loop variable always remains local to the list comprehension. It is never leaked outside.

We can verify this below:

```
>>> [... for loop_var in range(6)]
>>> print(loop_var)
NameError: name 'loop_var' is not defined
```

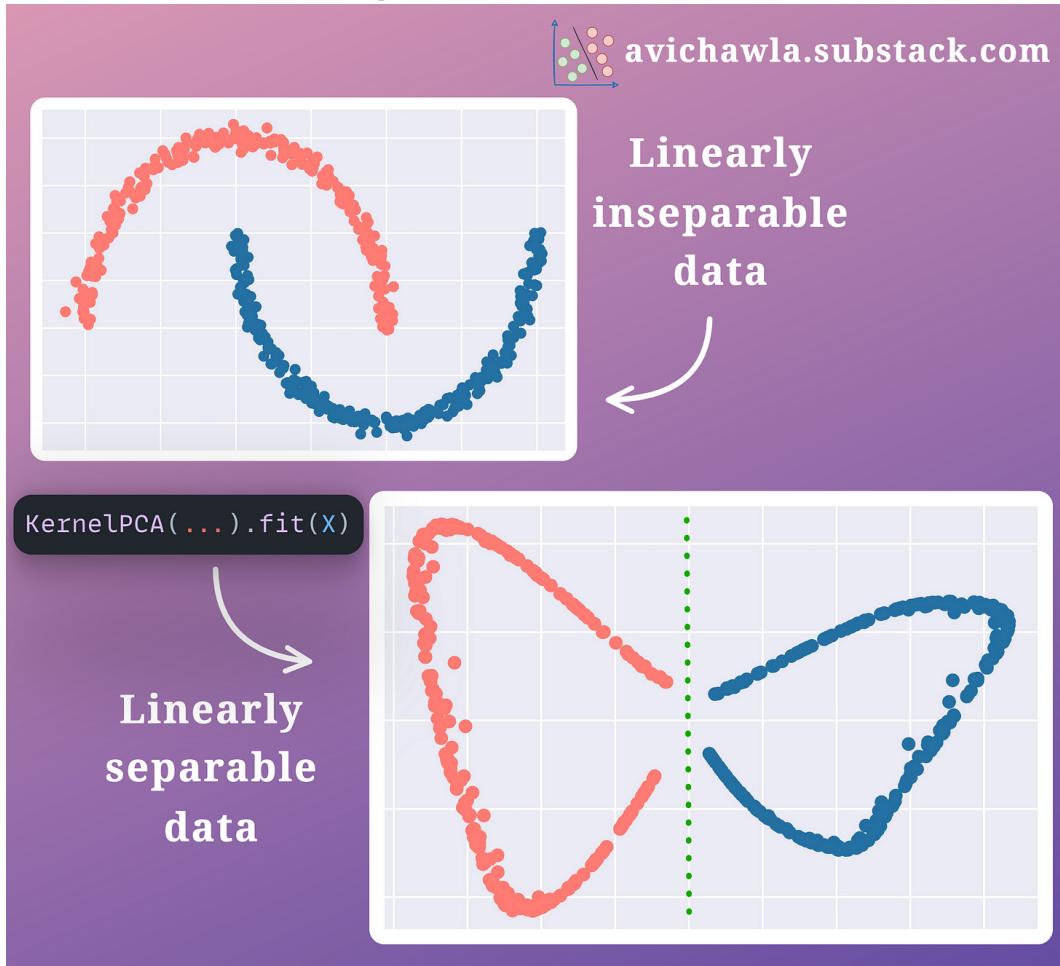
Error

That is why the existing variable (`a`), which was also used inside the list comprehension, remained unchanged. The list comprehension defined the loop variable (`a`) local to its scope.

Over to you: What are some other differences that you know of between for-loops and list comprehension?

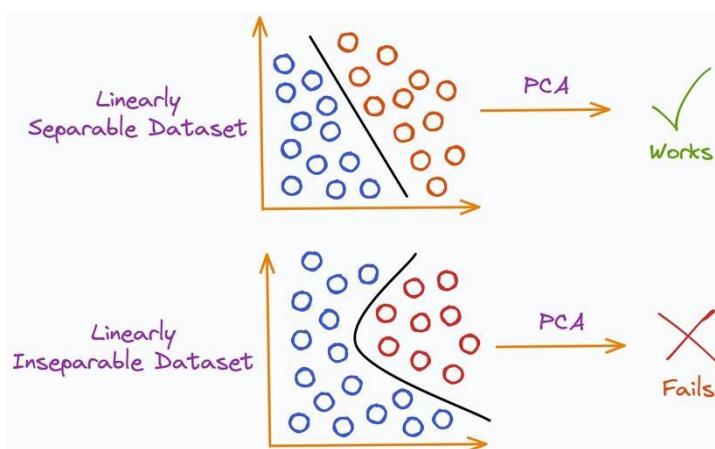


The Limitation of PCA Which Many Folks Often Ignore



Imagine you have a classification dataset. If you use PCA to reduce dimensions, it is inherently assumed that your data is linearly separable.

But it may not be the case always. Thus, PCA will fail in such cases.





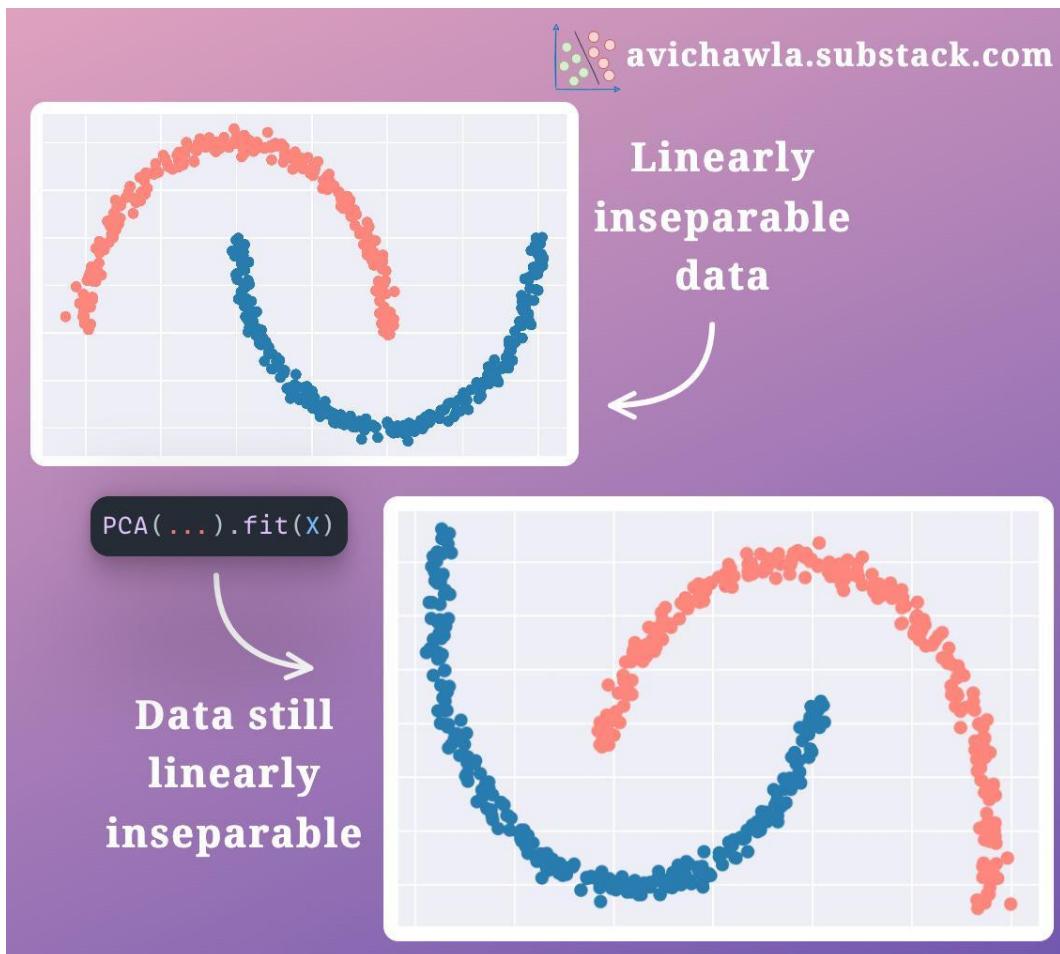
If you wish to read how PCA works, I would highly recommend reading one of my previous posts: [A Visual and Overly Simplified Guide to PCA](#).

To resolve this, we use the kernel trick (or the KernelPCA). The idea is to:

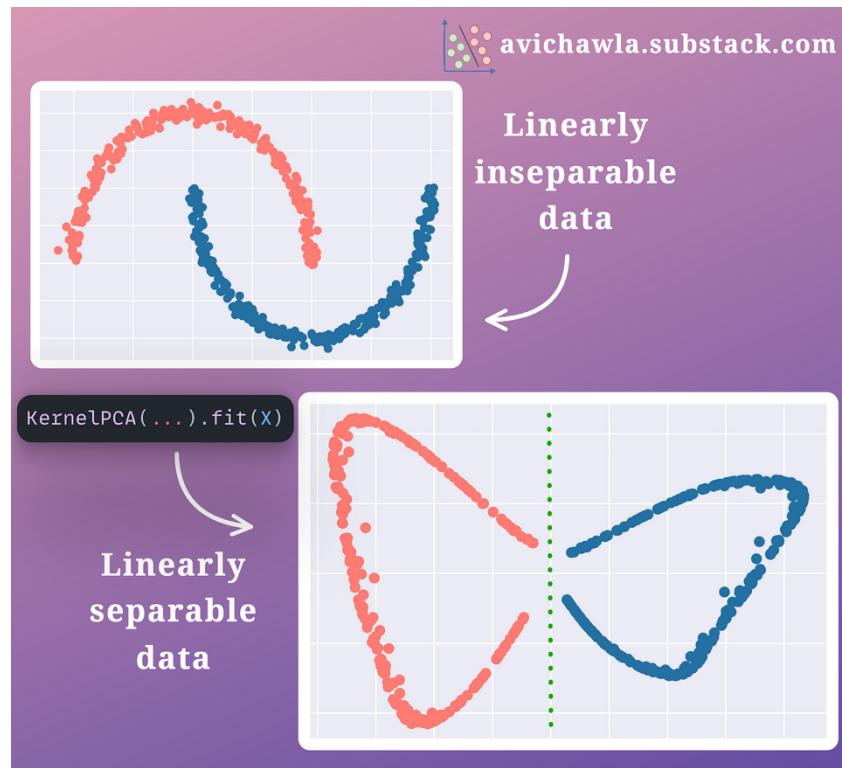
Project the data to another space using a kernel function, where the data becomes linearly separable.

Apply the standard PCA algorithm to the transformed data.

For instance, in the image below, the original data is linearly inseparable. Using PCA directly does not produce any desirable results.



But as mentioned above, KernelPCA first transforms the data to a linearly separable space and then applies PCA, resulting in a linearly separable dataset.



Sklearn provides a KernelPCA wrapper, supporting many popularly used kernel functions. You can find more details here: [Sklearn Docs](#).

Having said that, it is also worth noting that the run-time of PCA is cubic in relation to the number of dimensions of the data.

$$\text{Runtime} : O(nd^2 + d^3)$$

d : dimensions

n : samples

When we use a KernelPCA, typically, the original data (in n dimensions) is projected to a new higher dimensional space (in m dimensions; $m > n$). Therefore, it increases the overall run-time of PCA.



Magic Methods: An Underrated Gem of Python OOP



Magic Method	Syntax	Usage/Description
<code>__new__</code>	<code>__new__(cls, *args, **kwargs):</code>	Invoked before <code>__init__</code> to allocate memory to object
<code>__init__</code>	<code>__init__(self, *args, **kwargs):</code>	Invoked after <code>__new__</code> to initialise the object
<code>__str__</code>	<code>__str__(self):</code>	Invoked when <code>str(obj)</code> or <code>print(obj)</code> is used
<code>__int__</code>	<code>__int__(self):</code>	Invoked when <code>int(obj)</code> is used
<code>__len__</code>	<code>__len__(self):</code>	Invoked when <code>len(obj)</code> is used
<code>__call__</code>	<code>__call__(self, *args, **kwargs):</code>	Invoked when class object is called as a function: <code>obj()</code>
<code>__getitem__</code>	<code>__getitem__(self, key):</code>	Invoked when object is indexed: <code>obj[key]</code>
<code>__setitem__</code>	<code>__setitem__(self, key, value):</code>	Invoked when object is indexed and value is set: <code>obj[key]=value</code>
<code>__delitem__</code>	<code>__delitem__(self, key):</code>	Invoked when object's index is deleted: <code>del obj[key]</code>
<code>__contains__</code>	<code>__contains__(self, item):</code>	Invoked when the <code>in</code> operator is used: <code>item in obj</code>
<code>__bool__</code>	<code>__bool__(self):</code>	Invoked when object is used in boolean context: <code>if obj</code> or <code>bool(obj)</code>
<code>__iter__</code>	<code>__iter__(self):</code>	Invoked when object is iterated: <code>for x in obj</code>
<code>__eq__</code>	<code>__eq__(self, other):</code>	Invoked when <code>==</code> operator is used to compare two objects: <code>obj1 == obj2</code>
<code>__ne__</code>	<code>__ne__(self, other):</code>	Invoked when <code>!=</code> operator is used to compare two objects: <code>obj1 != obj2</code>
<code>__add__</code>	<code>__add__(self, other):</code>	Invoked when two objects are added: <code>obj1 + obj2</code>
<code>__mul__</code>	<code>__mul__(self, other):</code>	Invoked when two objects are multiplied: <code>obj1 * obj2</code>
<code>__abs__</code>	<code>__abs__(self):</code>	Invoked to compute absolute value of object: <code>abs(obj)</code>
<code>__neg__</code>	<code>__neg__(self):</code>	Invoked when unary operator <code>-</code> is used on an object: <code>-obj</code>
<code>__invert__</code>	<code>__invert__(self):</code>	Invoked when <code>~(tilde)</code> operator is used to invert an object: <code>~obj</code>

Magic Methods (also called **dunder methods**) are special methods defined inside a Python class' implementation.

*On a side note, the word “Dunder” is short for **Double Underscore**.*

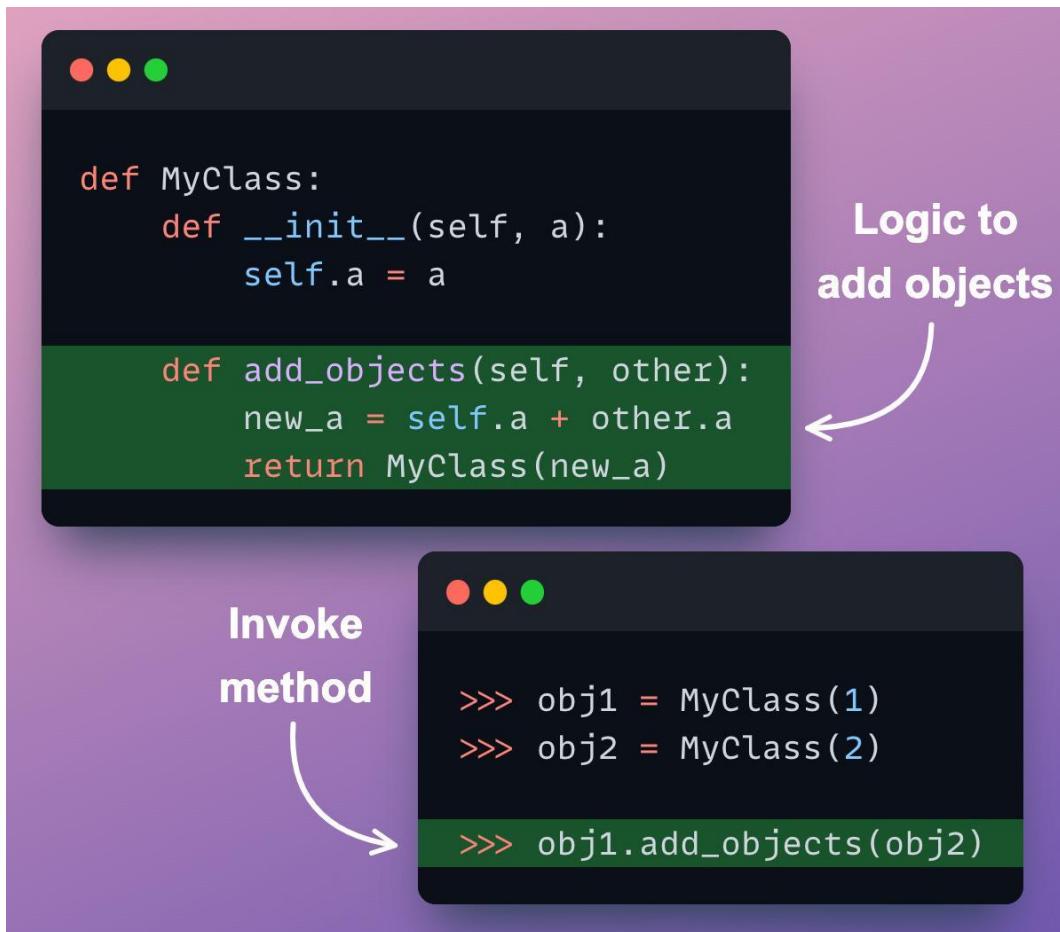
They are prefixed and suffixed with double underscores, such as `__len__`, `__str__`, and many more.



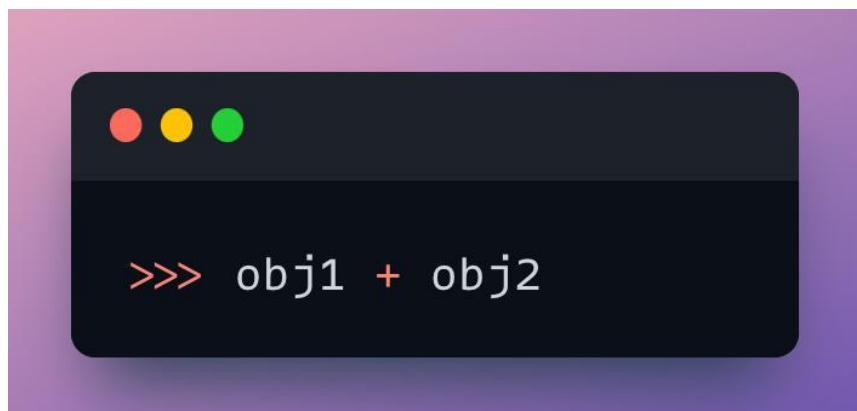
Magic Methods offer immense flexibility to define the behavior of class objects in certain scenarios.

For instance, say we want to define a custom behavior for adding two objects of our class (`obj1 + obj2`).

An obvious and straightforward way to do this is by defining a method, say `add_objects()`, and passing the two objects as its argument.



While the above approach will work, invoking a method explicitly for adding two objects isn't as elegant as using the `+` operator:





This is where magic methods come in. In the above example, implementing the `__add__` magic method will allow you to add the two objects using the `+` operator instead.

Thus, magic methods allow us to make our classes more intuitive and easier to work with.

As a result, awareness about them is extremely crucial for developing elegant, and intuitive pipelines.

The visual summarizes **~20** most commonly used magic methods in Python.

Over to you: What other magic methods will you include here? Which ones do you use the most? Let me know :)



The Taxonomy Of Regression Algorithms That Many Don't Bother To Remember

Regression Type	Description	Equation/Loss Function
Linear Regression	Simple Linear Regression	$\hat{y} = wx + b$ $Loss = \sum \frac{(y - \hat{y})^2}{n}$
	Polynomial Linear Regression	$\hat{y} = w_1x + w_2x^2 + \dots + b$ $Loss = \sum \frac{(y - \hat{y})^2}{n}$
	Multiple Linear Regression	$\hat{y} = w_1x_1 + w_2x_2 + \dots + b$ $Loss = \sum \frac{(y - \hat{y})^2}{n}$
Regularized Regression	Ridge Regression	$Loss = \sum \frac{(y - \hat{y})^2}{n} + \lambda \sum_{i=1}^n w_i^2$
	Lasso Regression	$Loss = \sum \frac{(y - \hat{y})^2}{n} + \lambda \sum_{i=1}^n w_i $
	Elastic Net	$Loss = \sum \frac{(y - \hat{y})^2}{n} + \lambda((1 - \alpha) * \sum_{L2} w_i^2 + \alpha * \sum_{L1} w_i)$
Categorical Probability	Logistic Regression	$P(X) = \frac{1}{1 + e^{-(w_1x_1 + w_2x_2 + \dots + b)}}$
	Multinomial Logistic Regression (or Softmax Regression)	$P(Y = k X) = \frac{e^{score_k}}{\sum_{j=1}^K e^{score_j}}$

Regression algorithms allow us to model the relationship between a dependent variable and one or more independent variables.

After estimating the parameters of a regression model, we can gain insight into how changes in one variable affect another.

Being widely used in data science, an awareness of their various forms is crucial to precisely convey which algorithm you are using.

Here are eight of the most standard regression algorithms described in a single line:

Linear Regression

- **Simple linear regression:** One independent (x) and one dependent (y) variable.



- **Polynomial Linear Regression:** Polynomial features and one dependent (y) variable.
- **Multiple Linear Regression:** Arbitrary features and one dependent (y) variable.

Regularized Regression

- **Lasso Regression:** Linear Regression with L1 Regularization.
- **Ridge Regression:** Linear Regression with L2 Regularization.
- **Elastic Net:** Linear Regression with BOTH L1 and L2 Regularization.

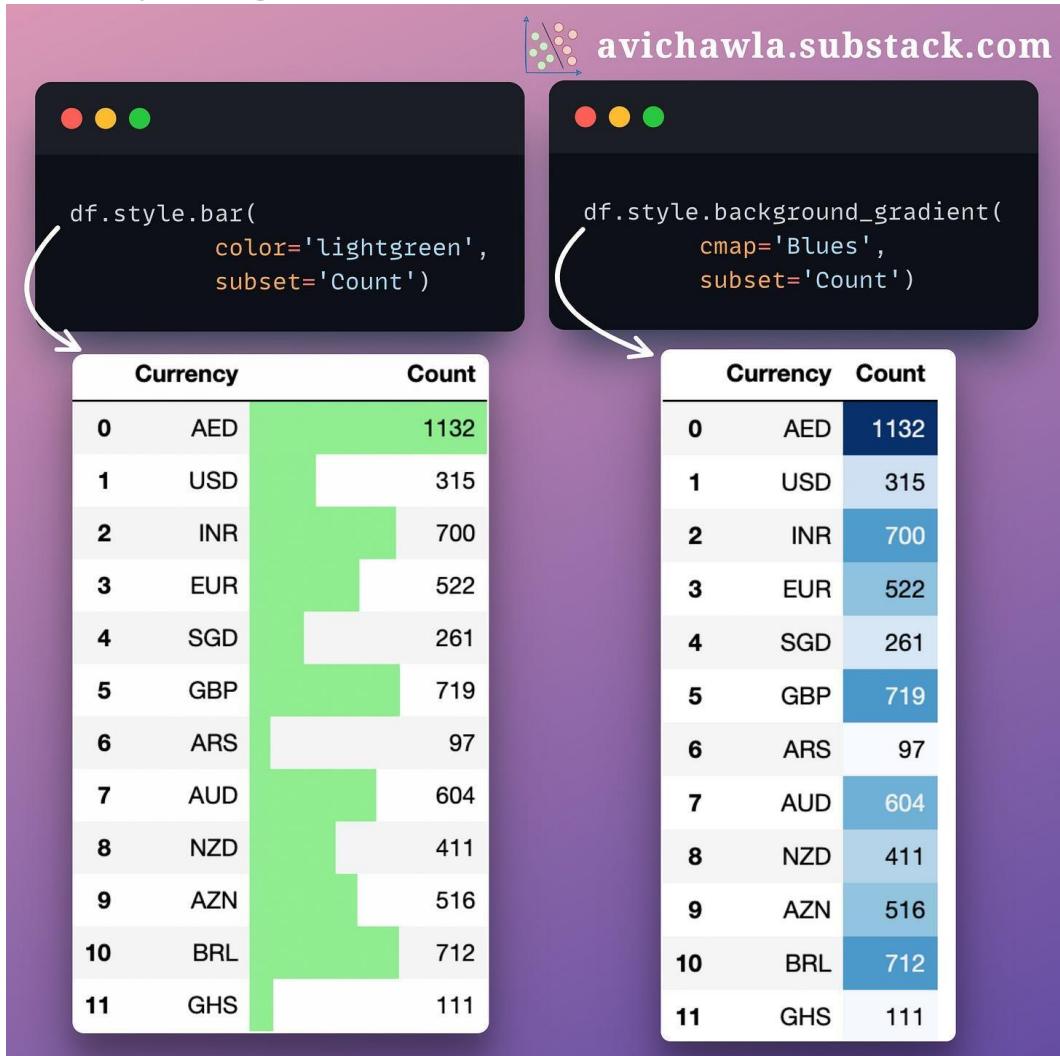
Categorical Probability Prediction

- **Logistic Regression:** Predict binary outcome probability.
- **Multinomial Logistic Regression (or Softmax Regression):** Predict multiple categorical probabilities.

Over to you: What other regressions algorithms will you include here?



A Highly Overlooked Approach To Analysing Pandas DataFrames



Instead of previewing raw DataFrames, styling can make data analysis much easier and faster. Here's how.

Jupyter is a web-based IDE. Anything you print is rendered using HTML and CSS.

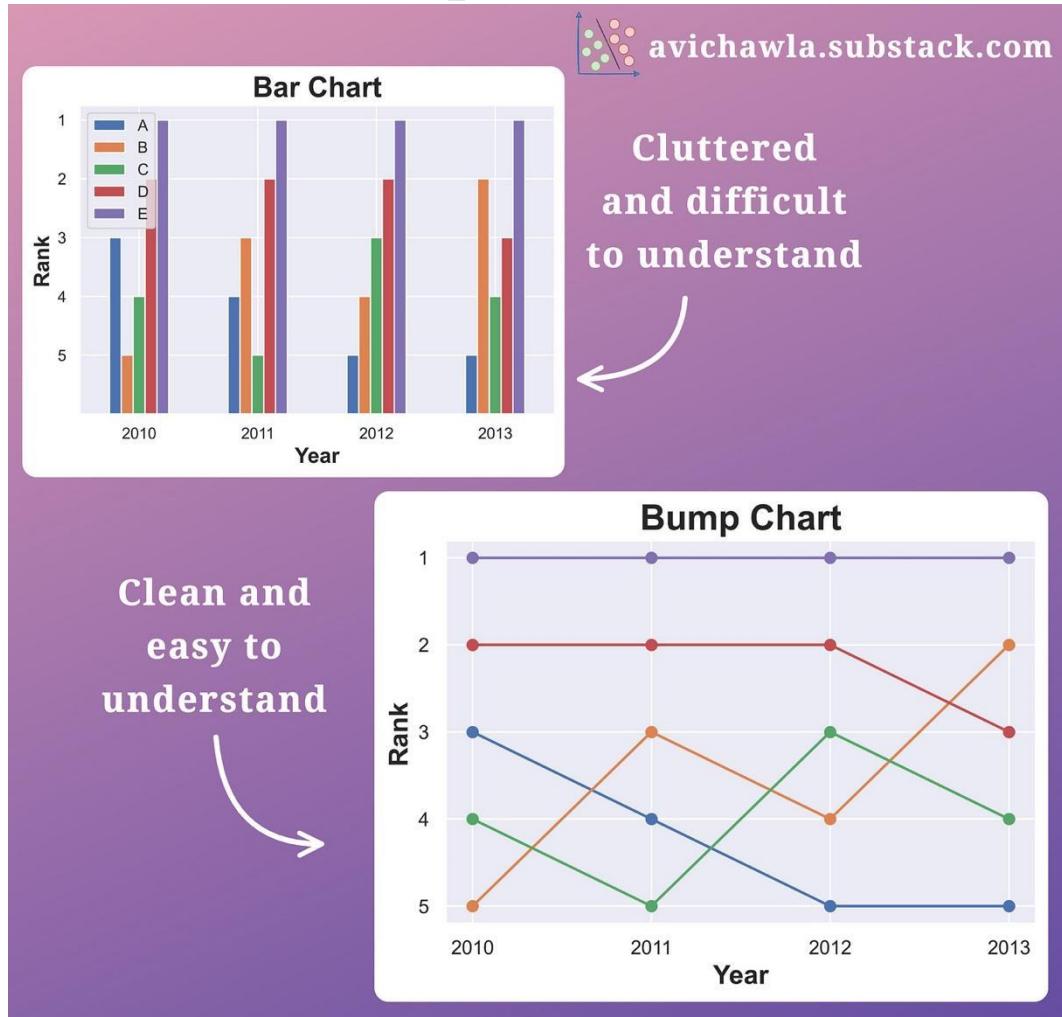
This means you can style your output in many different ways.

To style Pandas DataFrames, use its Styling API (**df.style**). As a result, the DataFrame is rendered with the specified styling.

Read more here: [Documentation](#).



Visualise The Change In Rank Over Time With Bump Charts



When visualizing the change in rank over time, using a bar chart may not be appropriate. Instead, try Bump Charts.

They are specifically used to visualize the rank of different items over time.

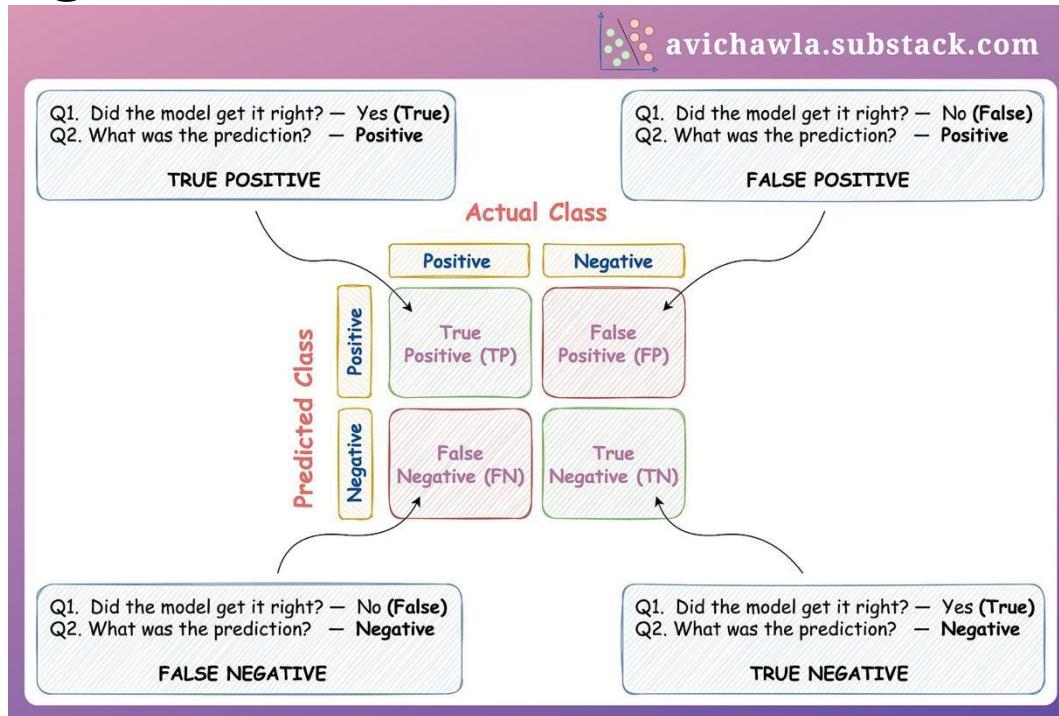
In contrast to the commonly used bar chart, they are clear, elegant, and easy to comprehend.

Over to you: What are some other bar chart alternatives to try in such cases? Let me know :)

Find the code for creating a bump chart in Python here: [Notebook](#).



Use This Simple Technique To Never Struggle With TP, TN, FP and FN Again



Do you often struggle to label model predictions as TP, TN, FP and FN, and comprehend them? If yes, here's a simple guide to help you out.

When labeling any prediction, ask yourself two questions:

Did the model get it right? **Answer: yes (or True) / no (or False).**

What was the predicted class? **Answer: Positive/Negative.**

Next, combine the above two answers to get the final label.

For instance, say the actual and predicted class were positive.

Did the model get it right? The answer is **yes (or TRUE).**

What was the predicted class? The answer is **POSITIVE.**

Final label: **TRUE POSITIVE.**

As an exercise, try labeling the following predictions. Consider the “Cat” class as “Positive” and the “Dog” class as “Negative”.



True Class	Predicted Class
	
	
	
	

Did the model get it right?	What was the predicted class?
-	-
-	-
-	-
-	-



The Most Common Misconception About Inplace Operations in Pandas



Method	Run-time	
	<i>inplace=False</i>	<i>inplace=True</i>
<code>df.replace()</code>	140 μs	244 μs (Slow)
<code>df.sort_values()</code>	374 μs	450 μs (Slow)
<code>df.reset_index()</code>	35 μs	10 μs (Fast)
<code>df.drop()</code>	200 μs	262 μs (Slow)
<code>df.fillna()</code>	90 μs	222 μs (Slow)
<code>df.dropna()</code>	750 μs	1088 μs (Slow)
<code>df.drop_duplicates()</code>	856 μs	1058 μs (Slow)
<code>df.rename()</code>	151 μs	152 μs (Equal)

Pandas users often modify a DataFrame inplace expecting better performance. Yet, it may not always be efficient. Here's why.

The image compares the run-time of inplace and non-in-place operations. In most cases, inplace operations are slow.

Why?

Contrary to common belief, most inplace operations DO NOT prevent the creation of a new copy. It is just that inplace assigns the copy back to the same address.



But during this assignment, Pandas performs some extra checks (SettingWithCopy) to ensure that the DataFrame is being modified correctly. This, at times, can be an expensive operation.

Yet, in general, there is no guarantee that an inplace operation is faster.

What's more, inplace operations do not allow chaining multiple operations, such as this:

The screenshot shows two code snippets side-by-side. The top snippet, titled "Method chaining", contains the following code:

```
df.reset_index().fillna(0).drop_duplicates()
```

The bottom snippet, titled "No chaining with Inplace", contains the following code:

```
df.reset_index(inplace = True)  
df.fillna(0, inplace = True)  
df.drop_duplicates(inplace = True)
```

A red arrow points from the "No chaining with Inplace" title towards the "df.fillna(0, inplace = True)" line in the bottom code block.



Build Elegant Web Apps Right From Jupyter Notebook with Mercury

The image shows a comparison between a Jupyter Notebook cell and the resulting web application generated by Mercury.

Jupyter Notebook Cell:

```
import mercury as mr
app = mr.App(title="Mercury App", description="Mercury demo")
Linear Data App
name = mr.Text(label="What is your name?", value="Avi")
points = mr.Slider(label="Number of points", value=75, min=50, max=100)
color = mr.Select(label="Select color", value="green", choices=['red', 'green', 'blue'])
mr.Md(f"\#\# Welcome {name.value}")
Welcome Avi.
x = np.linspace(0, 10, points.value)
y = 3*x + 5 + 10*random(points.value)
_ = plt.scatter(x, y, color=color.value)
_ = plt.title("Linear Data Plot")
Linear Data Plot
```

Mercury Web App:

MERCURY

Mercury App

What is your name? Avi

Number of points 75

Select color blue

Download Share

Linear Data

Welcome Avi.

Linear Data Plot

A white arrow points from the Jupyter cell to the "Mercury web app" section, and another white arrow points from the "Mercury web app" section to the generated web interface.

Exploring and sharing data insights using Jupyter is quite common for data folks. Yet, an interactive app is better for those who don't care about your code and are interested in your results.

While creating presentations is possible, it can be time-consuming. Also, one has to leave the comfort of a Jupyter Notebook.

Instead, try Mercury. It's an open-source tool that converts your jupyter notebook to a web app in no time. Thus, you can create the web app without leaving the notebook.

A quick demo is shown below:



The screenshot shows the Mercury app interface. On the left, a Jupyter Notebook cell (In [1]) contains code to import mercury, numpy, seaborn, and matplotlib, and to set sns. A second cell (In [2]) creates a Mercury app titled "Mercury App" with a description "Mercury demo". A third cell (In [3]) defines a "Linear Data App" with a text input for name ("What is your name?") and a dropdown for color ("Select color"). A fourth cell (In [4]) displays a welcome message "Welcome Avi." and a scatter plot titled "Linear Data Plot" with blue points. On the right, a "MERCURY" header is followed by a "Mercury App" section where "What is your name?" is set to "Avi" and "Number of points" is set to 75. Below this is a "Linear Data App" section with a welcome message "Welcome Avi.", a slider for "Number of points" (set to 75), a dropdown for "Select color" (set to "blue"), and a scatter plot titled "Linear Data Plot" showing blue points.

What's more, all updates to the Jupyter Notebook are instantly reflected in the Mercury app.

In contrast to the widely-adopted streamlit, web apps created with Mercury can be:

Exported as PDF/HTML.

Showcased as a live presentation.

Secured with authentication to restrict access.



Become A Bilingual Data Scientist With These Pandas to SQL Translations



avichawla.substack.com

Operation	Pandas	SQL
Read CSV	<code>pd.read_csv(file)</code>	<code>LOAD DATA INFILE 'data.csv' INTO TABLE table FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 ROWS;</code>
Print first 10 (or k) rows	<code>df.head(10)</code>	<code>SELECT * FROM table LIMIT 10;</code>
Dimensions	<code>df.shape</code>	<code>SELECT count(*) FROM table;</code>
Datatype	<code>df.dtypes</code>	<code>DESCRIBE table;</code>
Filter Data	<code>df[df.column>10]</code>	<code>SELECT * FROM table where column>10;</code>
Select column(s)	<code>df.column</code>	<code>SELECT column FROM table;</code>
Sort	<code>df.sort_values("column")</code>	<code>SELECT * FROM table ORDER BY column;</code>
Fill NaN	<code>df.column.fillna(0)</code>	<code>UPDATE table SET column=0 WHERE column IS NULL;</code>
Join	<code>pd.merge(df1, df2, on ="col", how = "inner")</code>	<code>SELECT * FROM table1 JOIN table2 ON (table1.col = table2.col);</code>
Concatenate	<code>pd.concat((df1, df2))</code>	<code>SELECT * FROM table1 UNION ALL table2;</code>
Group	<code>df.groupby("column"). agg_col.mean()</code>	<code>SELECT column, avg(agg_col) FROM table GROUP BY column;</code>
Unique values	<code>df.column.unique()</code>	<code>SELECT DISTINCT column FROM table;</code>
Rename column	<code>df.rename(columns = {"old_name": "new_name"})</code>	<code>ALTER TABLE table RENAME COLUMN old_name TO new_name;</code>
Delete column	<code>df.drop(columns = ["column"])</code>	<code>ALTER TABLE table DROP COLUMN column;</code>

SQL and Pandas are both powerful tools for data scientists to work with data.

Together, SQL and Pandas can be used to clean, transform, and analyze large datasets, and to create complex data pipelines and models.



Thus, proficiency in both frameworks can be extremely valuable to data scientists.

This visual depicts a few common operations in Pandas and their corresponding translations in SQL.

I have a detailed blog on Pandas to SQL translations with many more examples.

Read it here: [Pandas to SQL blog](#).

Over to you: What other Pandas to SQL translations will you include here?



A Lesser-Known Feature of Sklearn To Train Models on Large Datasets

Failed due to
memory
constraints

```
from sklearn.linear_model
import SGDClassifier

clf = SGDClassifier(...)

clf.fit(X, y)
```

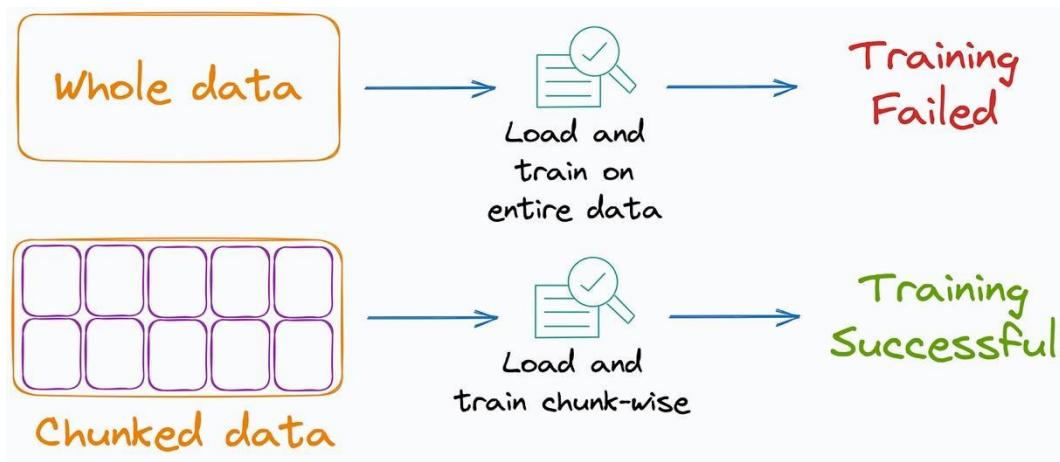
```
# 1. Load data in chunks (say, 1000 rows at once).
data = pd.read_csv("data.csv", chunksize = 1000)

# 2. Train from mini-batches using partial_fit.
for batch in data:
    clf.partial_fit(batch["X"],
                    batch["y"],
                    classes = [0, 1, 2])
```

It is difficult to train models with sklearn when you have plenty of data. This may often raise memory errors as the entire data is loaded in memory. But here's what can help.

Sklearn implements the **partial_fit** API for various algorithms, which offers incremental learning.

As the name suggests, the model can learn incrementally from a mini-batch of instances. This prevents limited memory constraints as only a few instances are loaded in memory at once.



As shown in the main image, `clf.fit(X, y)` takes the entire data, and thus, it may raise memory errors. But, loading chunks of data and invoking the `clf.partial_fit()` method prevents this and offers seamless training.

Also, remember that while using the **partial_fit** API, a mini-batch may not have instances of all classes (especially the first mini-batch). Thus, the model will be unable to cope with new/unseen classes in subsequent mini-batches. Therefore, you should pass a list of all possible classes in the `classes` parameter.

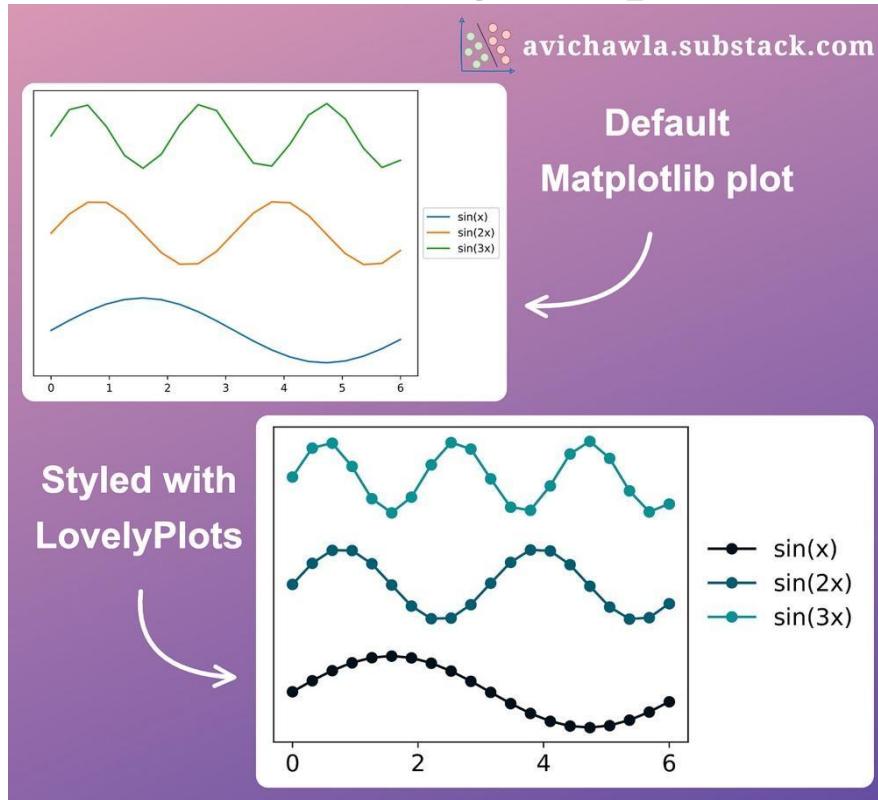
Having said that, it is also worth noting that not all sklearn estimators implement the **partial_fit** API. Here's the list:

- Classification
 - `sklearn.naive_bayes.MultinomialNB`
 - `sklearn.naive_bayes.BernoulliNB`
 - `sklearn.linear_model.Perceptron`
 - `sklearn.linear_model.SGDClassifier`
 - `sklearn.linear_model.PassiveAggressiveClassifier`
- Regression
 - `sklearn.linear_model.SGDRegressor`
 - `sklearn.linear_model.PassiveAggressiveRegressor`
- Clustering
 - `sklearn.cluster.MiniBatchKMeans`
- Decomposition / feature Extraction
 - `sklearn.decomposition.MiniBatchDictionaryLearning`
 - `sklearn.cluster.MiniBatchKMeans`

Yet, it is surely worth exploring to see if you can benefit from it :)



A Simple One-Liner to Create Professional Looking Matplotlib Plots



The default styling of matplotlib plots appears pretty basic at times. Here's how you can make them appealing.

To create professional-looking plots for presentations, reports, or scientific papers, try LovelyPlots.

It provides many style sheets to improve their default appearance, by simply adding just one line of code.

To install LovelyPlots, run the following command:

```
pip install LovelyPlots
```

Next, import the matplotlib library, and change the style as follows: (You don't have to import LovelyPlots anywhere)

```
import matplotlib.pyplot as plt  
plt.style.use(style) ## change to the style provided by LovelyPlots
```

Print the list of all possible styles as follows:

```
plt.style.available
```

Get Started: [LovelyPlots Repository](https://github.com/avichawla/LovelyPlots).



DailyDoseofDS.com



Avoid This Costly Mistake When Indexing A DataFrame

df.shape

(32768000, 9)



First column then row

```
%timeit df["col"]["row"]
```

2.96 μ s ± 7.17 ns per loop

Selecting a column first is over 15x faster

First row then column

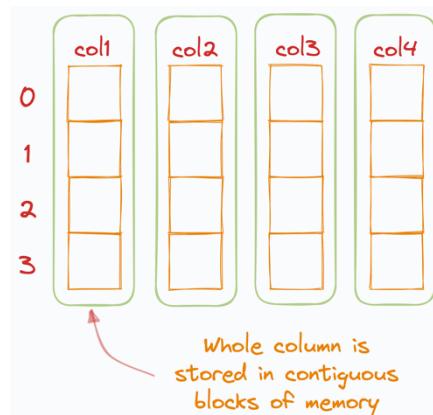
```
%timeit df.loc["row"]["col"]
```

45.4 μ s ± 384 ns per loop

When indexing a dataframe, choosing whether to select a column first or slice a row first is pretty important from a run-time perspective.

As shown above, selecting the column first is over **15 times** faster than slicing the row first. Why?

As I have talked before, Pandas DataFrame is a column-major data structure. Thus, consecutive elements in a column are stored next to each other in memory.

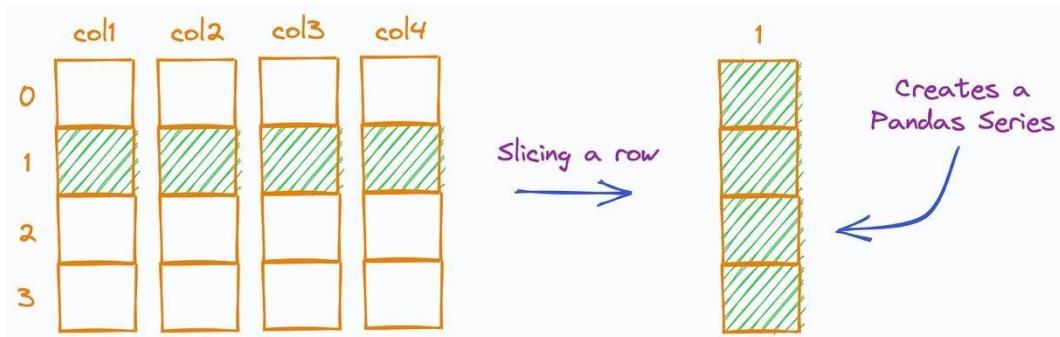




As processors are efficient with contiguous blocks of memory, accessing a column is much faster than accessing a row (read more about this in one of my previous posts [here](#)).

But when you slice a row first, each row is retrieved by accessing non-contiguous blocks of memory, thereby making it slow.

Also, once all the elements of a row are gathered, Pandas converts them to a Series, which is another overhead.



We can verify this conversion below:

A screenshot of a Jupyter Notebook cell. The code shows:

```
>>> df
   A   B
0  1  2
1  3  4

>>> df.loc[0]
A    1
B    2
Name: 0, dtype: int64

>>> type(df.loc[0])
pandas.core.series.Series
```

A white arrow points from the text "Creates a Pandas Series" in the previous diagram to the output of the `type(df.loc[0])` command in the screenshot, which shows `pandas.core.series.Series`.

Instead, when you select a column first, elements are retrieved by accessing contiguous blocks of memory, which is way faster. Also, a column is inherently a Pandas Series. Thus, there is no conversion overhead involved like above.



```
>>> df
   A   B
0  1  2
1  3  4

>>> df["A"]
0    1
1    3
Name: A, dtype: int64

>>> type(df["A"])
pandas.core.series.Series
```

Series object

Overall, by accessing the column first, we avoid accessing non-contiguous memory access, which does happen when we access the row first.

This makes selecting the column first faster than slicing a row first in indexing operations.

If you are confused about what selecting, indexing, slicing, and filtering mean, here's what you should read next:

<https://avichawla.substack.com/p/are-you-sure-you-are-using-the-correct>.



9 Command Line Flags To Run Python Scripts More Flexibly

Python Command Line Flags

Python Flag	Description	Usage
<code>python -c</code>	Run a single python command	<code>python -c "print('Hello')"</code>
<code>python -i</code>	Run interactive Python shell after running a script	<code>python -i script.py</code>
<code>python -O</code>	Ignore assert statements	<code>python -O script.py</code>
<code>python -OO</code>	Ignore assert statements and docstrings	<code>python -OO script.py</code>
<code>python -W</code>	Ignore warnings	<code>python -W script.py</code>
<code>python -m</code>	Run a module as a script	<code>python -m my_package.my_module</code>
<code>python -v</code>	Enable verbose mode. Prints more information about what interpreter is doing	<code>python -v script.py</code>
<code>python -x</code>	Ignore the first line of the script (often the shebang line)	<code>python -x script.py</code>
<code>python -E</code>	Ignore all Python Environment variables	<code>python -E script.py</code>



When invoking a Python script, you can specify various options/flags. They are used to modify the behavior of the Python interpreter when it runs a script or module.

Here are 9 of the most commonly used options:

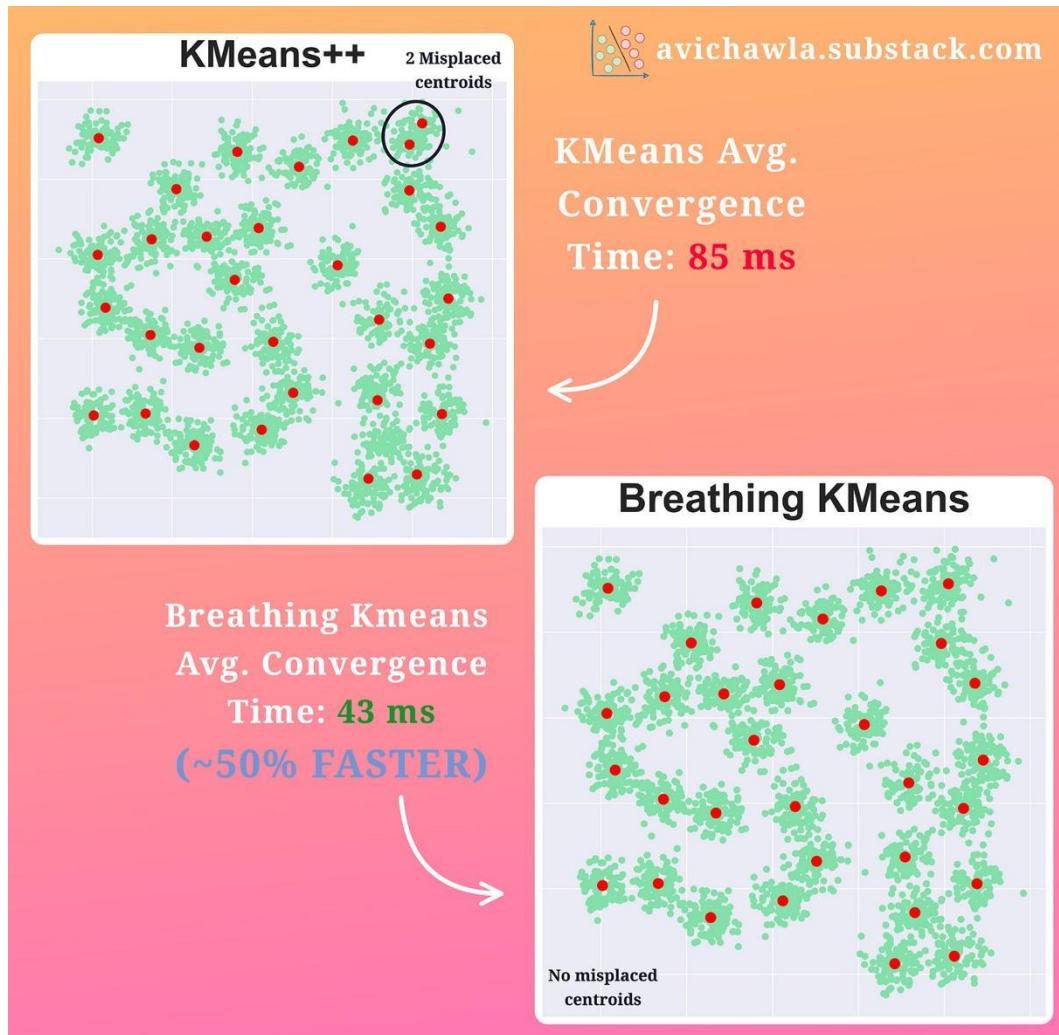


- ◆ **python -c:** Run a single Python command. Useful for running simple one-liners or testing code snippets.
- ◆ **python -i:** Run the script as usual and enter the interactive mode instead of terminating the program. Useful for debugging as you can interact with objects created during the program.
- ◆ **python -O:** Ignore assert statements (This is alphabet 'O'). Useful for optimizing code by removing debugging code.
- ◆ **python -OO:** Ignore assert statements and discard docstrings. Useful for further optimizing code by removing documentation strings.
- ◆ **python -W:** Ignore all warnings. Useful for turning off warnings temporarily and focusing on development.
- ◆ **python -m:** Run a module as a script.
- ◆ **python -v:** Enter verbose mode. Useful for printing extra information during program execution.
- ◆ **python -x:** Skip the first line. Useful for removing shebang lines or other comments at the start of a script.
- ◆ **python -E:** ignore all Python environment variables. Useful for ensuring a consistent program behavior by ignoring environment variables that may affect program execution.

Which ones have I missed? Let me know :)



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:

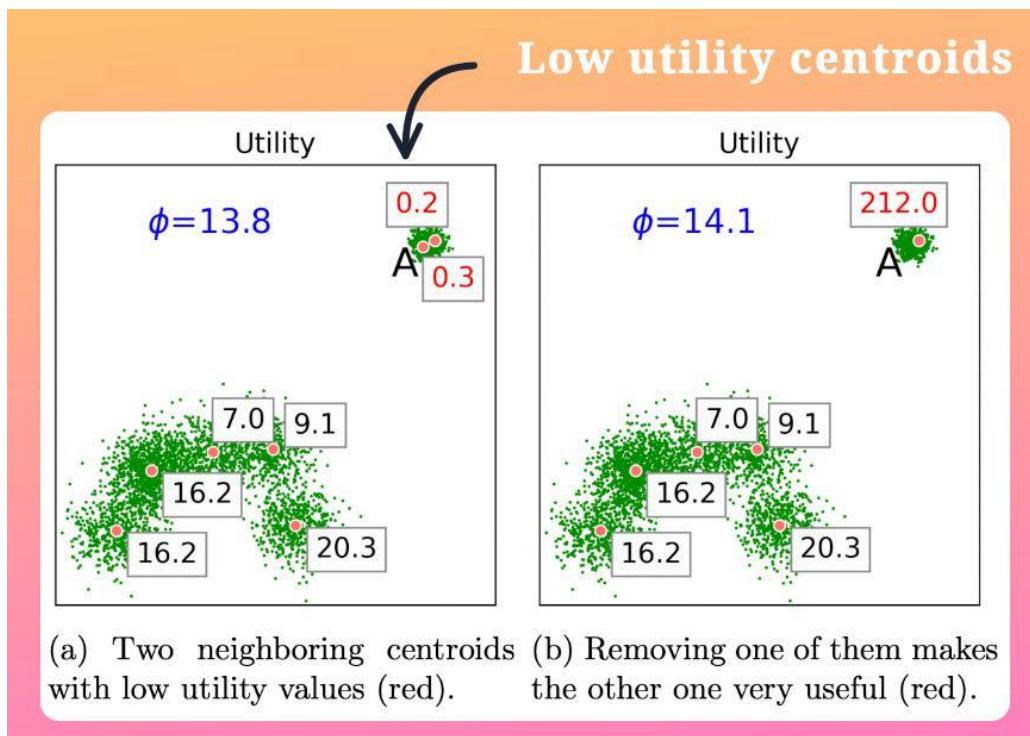
- **Step 1:** Initialise k centroids and run KMeans without repeating. In other words, don't re-run it with different initializations. Just run it once.
- **Step 2 — Breathe in step:** Add m new centroids and run KMeans with $(k+m)$ centroids without repeating.



- **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 $m=0$.

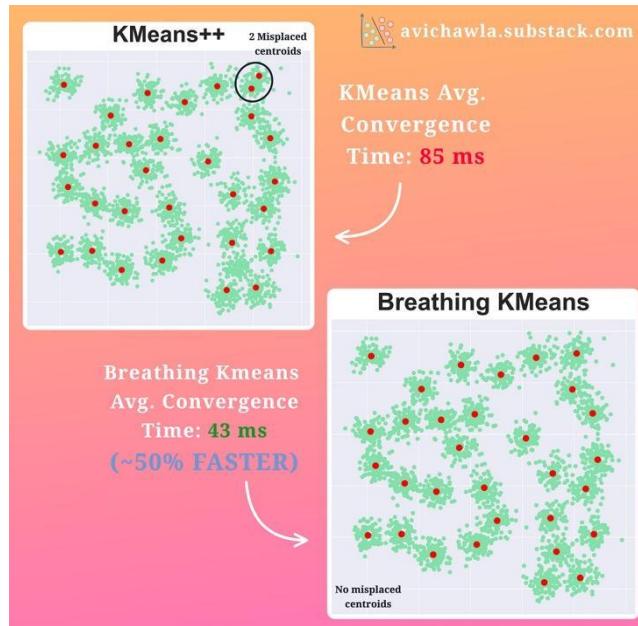
Breathe in step inserts new centroids close to the centroids with the largest errors. A centroid's error is the sum of the squared distance of points under that centroid.

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)

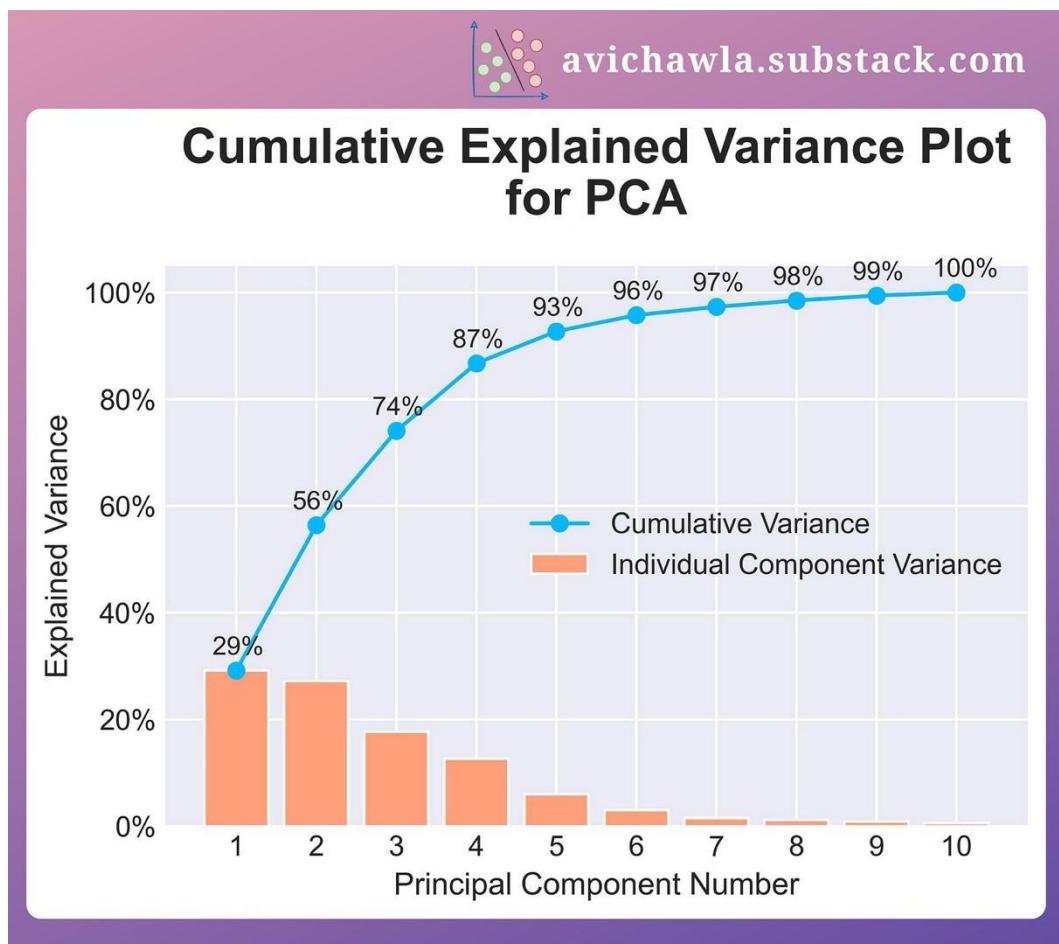
# run the algorithm
bkm.fit(X)
```

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](#).



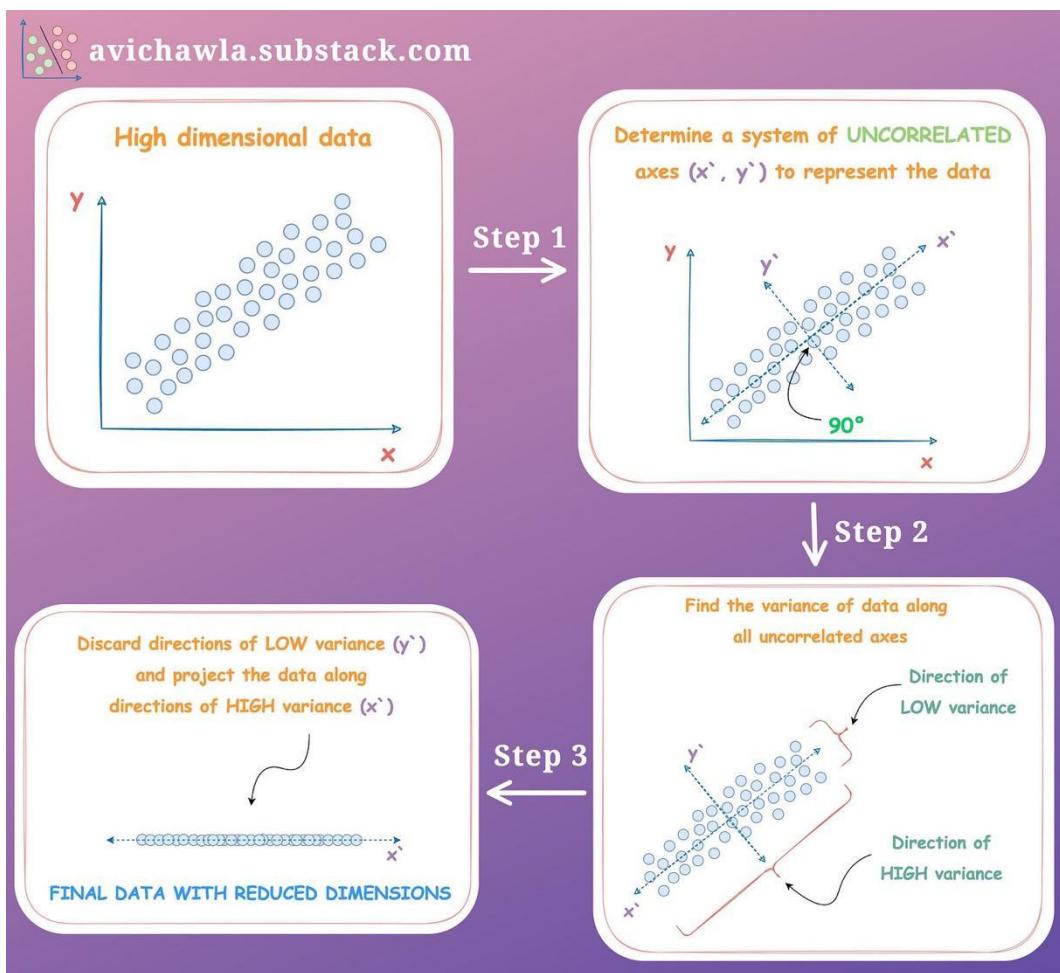
How Many Dimensions Should You Reduce Your Data To When Using PCA?



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.

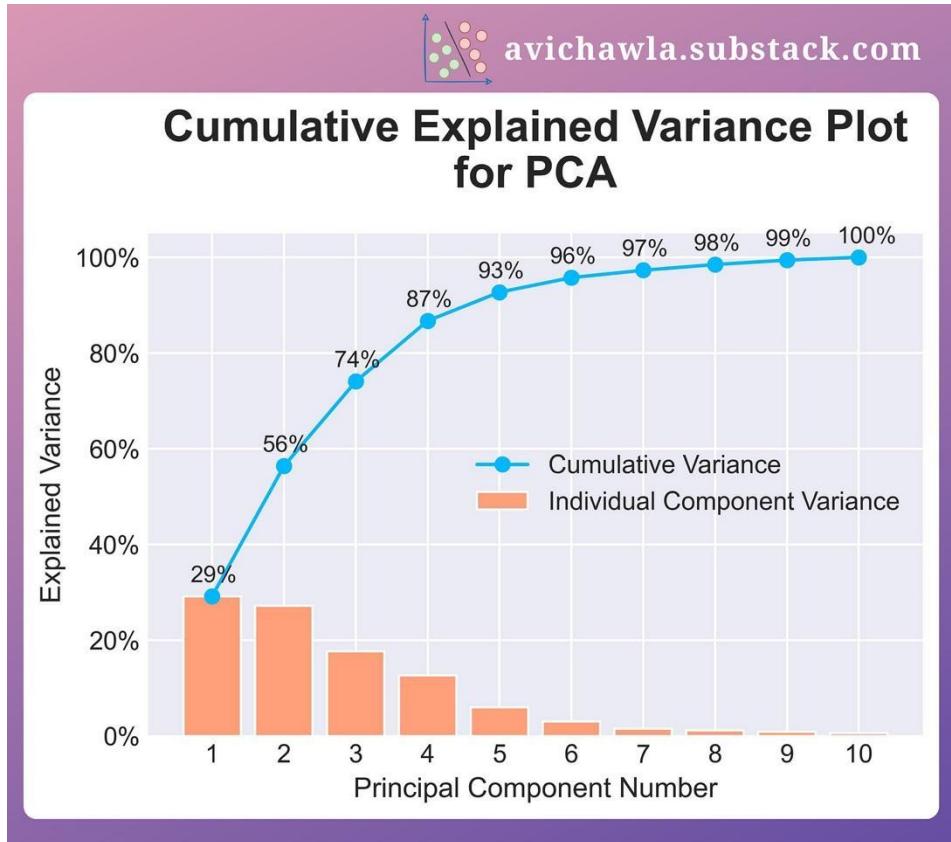
Step 3. Discard the axes with low variance. How many dimensions to discard (or keep) is a hyperparameter, which we will discuss below. Project the data along the retained axes.



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 interface with a Python code cell at the top:

```
In [1]: import mitosheet  
mitosheet.sheet(analyses_to_replay="id-utbdzhhmvd")
```

The interface includes a toolbar with various data manipulation tools like Undo, Redo, Clear, Import, Export, Add Col, Del Col, Dtype, Less, More, Number, Pivot, Graph, and AI. A green arrow points to the AI button in the toolbar. Below the toolbar is a data table titled "City | City". The table has columns: Name (str), Company (str), City (str), Salary (float), Status (str), and Rating (float). The data consists of 14 rows of employee information. At the bottom left is a dropdown menu for "employee_dataset", and at the bottom right is a note "(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-utbdzhmhvd");
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:

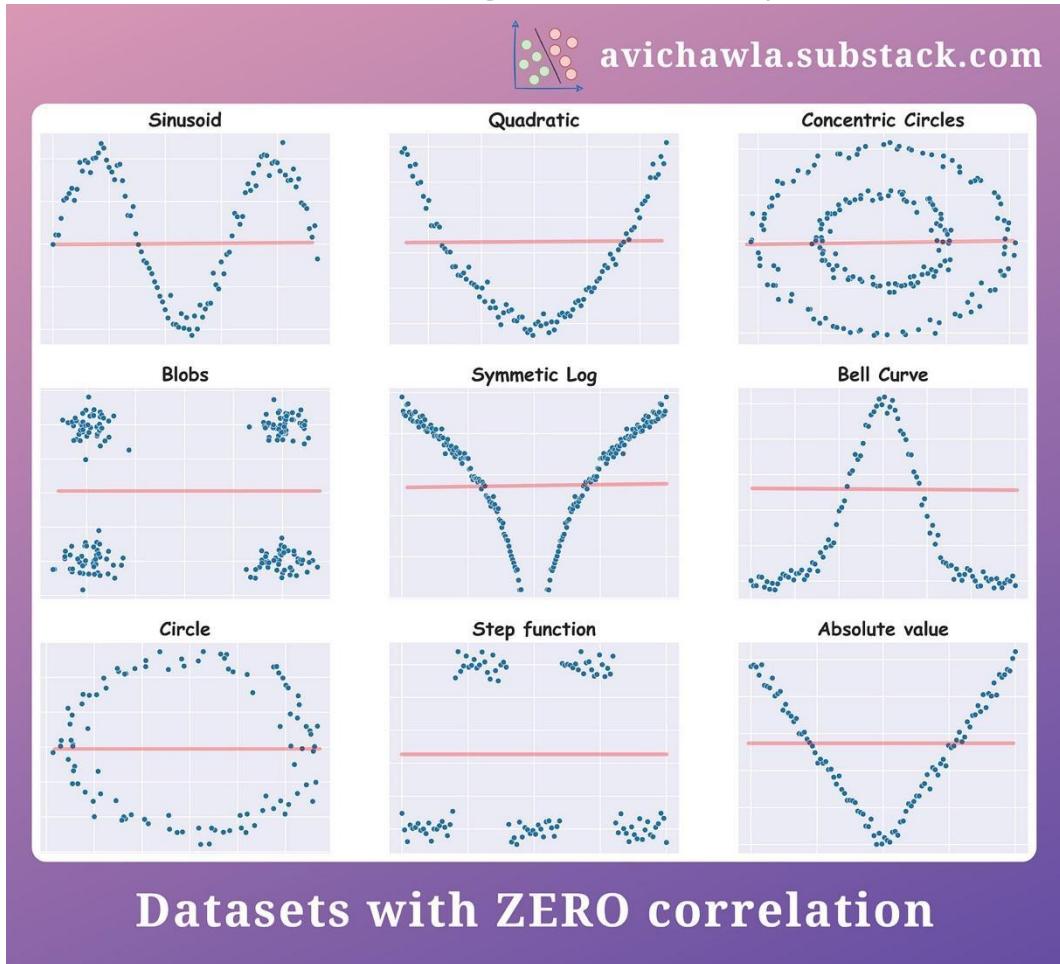
```
python -m pip install mitosheet
```

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



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 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

The diagram illustrates the behavior of custom Python objects in boolean contexts. On the left, a file named `without_bool.py` contains a `Cart` class with an `__init__` method but no `__bool__` method. An `if` statement checks if an instance of `Cart` is True, which it is, because by default, any object is considered True in a boolean context. The output is "Cart Not Empty". On the right, a file named `with_bool.py` contains the same `Cart` class, but it includes an `__bool__` method that returns `len(self.items) > 0`. This overrides the default behavior. When an instance of `Cart` is checked in a boolean context, it now evaluates to False because its length is zero. The output is "Cart Empty". Arrows point from the text "Object of custom class evaluated to True by default" to the left code and from "Object evaluated to False" to the right code.

```
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
```

```
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 of custom class evaluated to True by default

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.

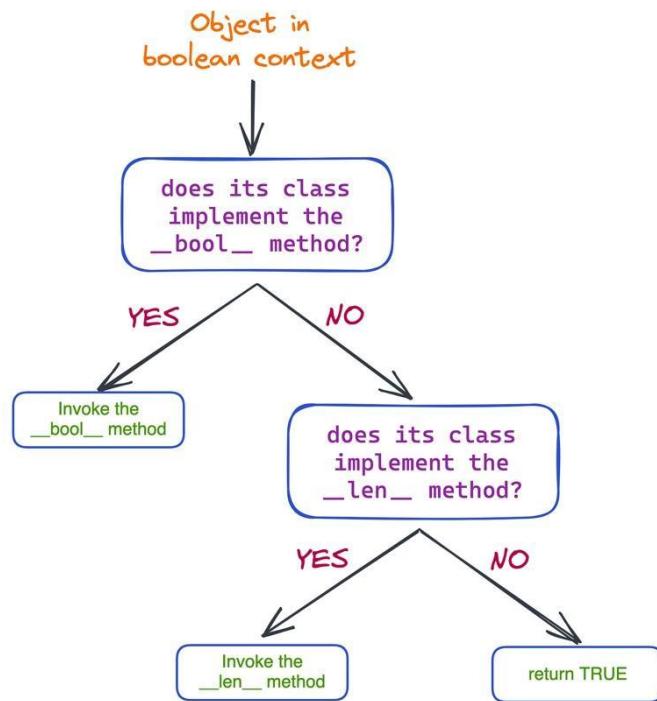
The `__bool__` dunder method is used to define the behavior of an object when used in a boolean context. As a result, you can specify explicit conditions to determine the truthiness of an object.

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:

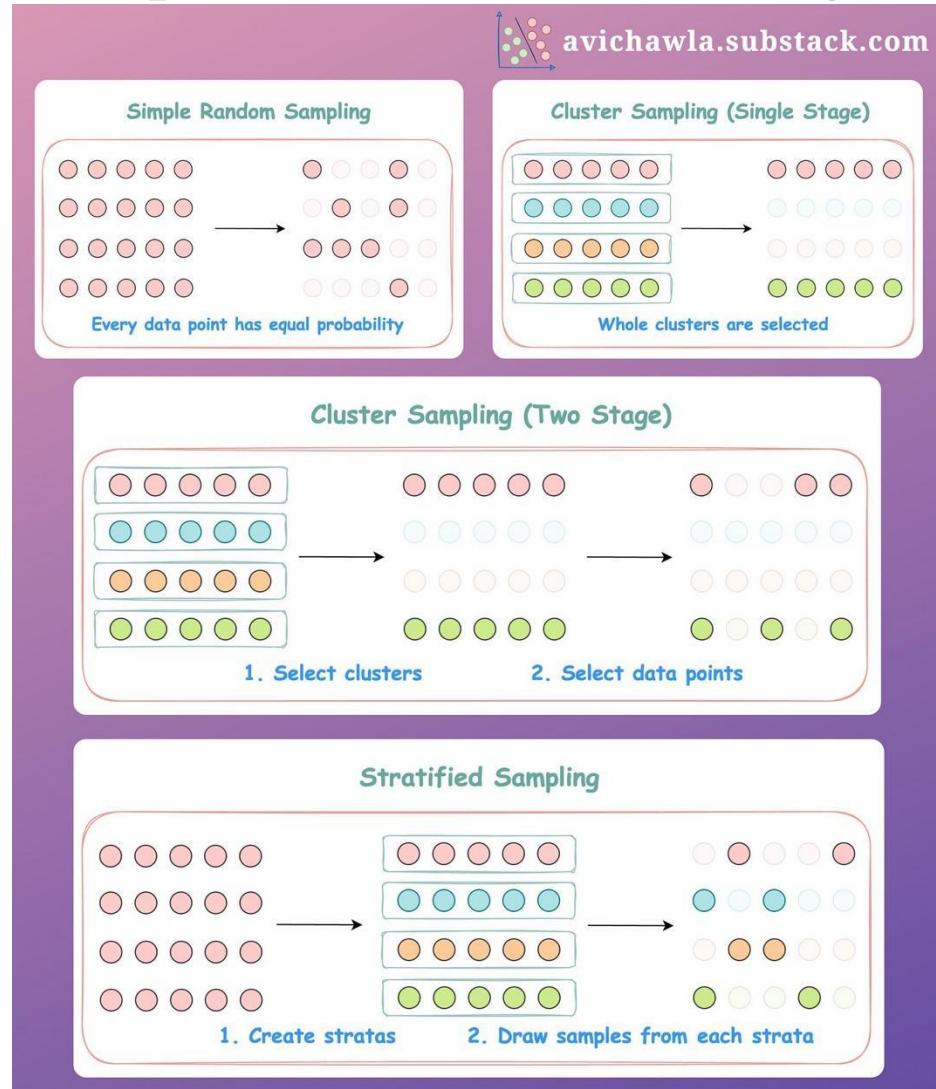


First, Python checks for the `__bool__` method in its class implementation. If found, it is invoked. If not, Python checks for the `__len__` method. If found, `__len__` is invoked. Otherwise, Python returns True.

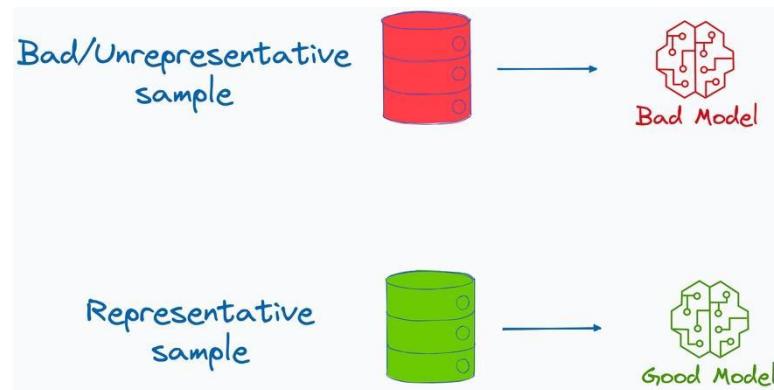
This explains the default behavior of objects instantiated from a custom class. As the `Cart` class implemented neither the `__bool__` method nor the `__len__` method, the `cart` object was evaluated to True.



A Visual Guide To Sampling Techniques in Machine Learning



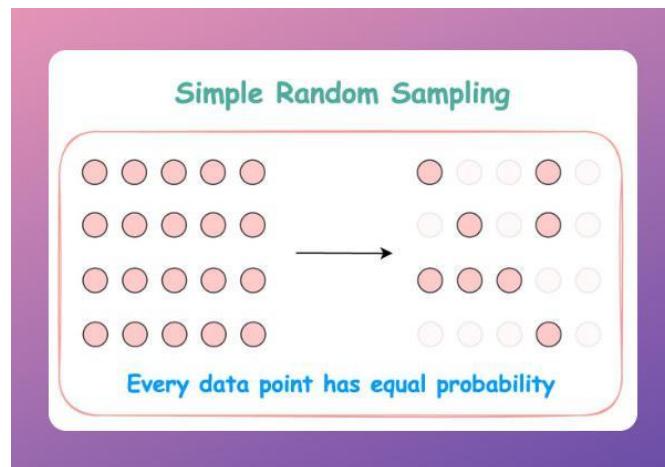
When you are dealing with large amounts of data, it is often preferred to draw a relatively smaller sample and train a model. But any mistakes can adversely affect the accuracy of your model.



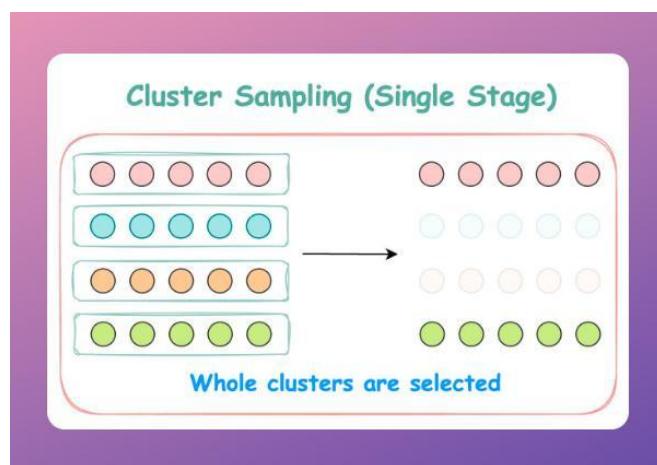
This makes sampling a critical aspect of training ML models.

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.

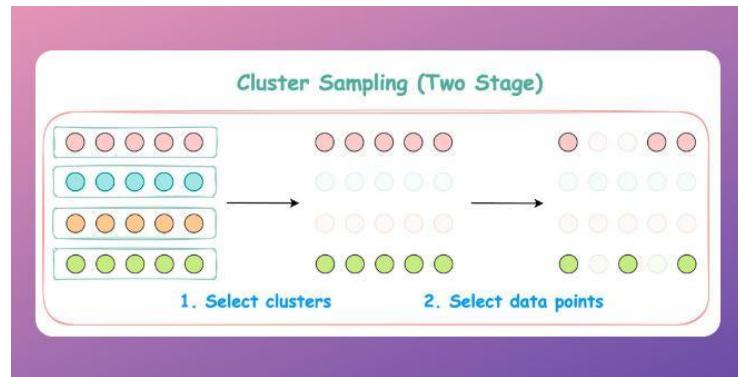


- ◆ **Cluster sampling (single-stage):** Divide the data into clusters and select a few entire clusters.



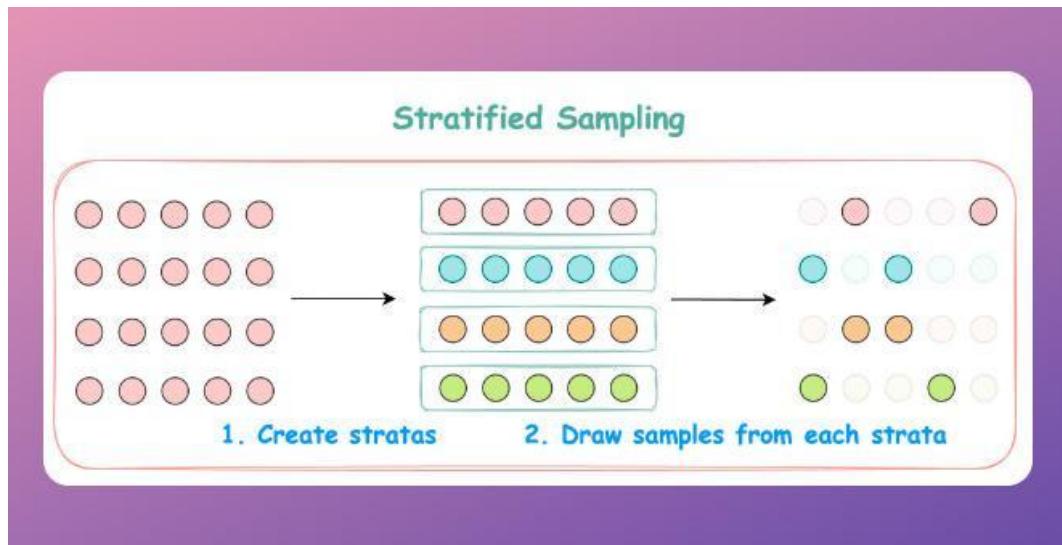


- ◆ **Cluster sampling (two-stage):** Divide the data into clusters, select a few clusters, and choose points from them randomly.





- ◆ **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

avichawla.substack.com

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:

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

```
>>> 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)
```

A curly brace on the left side of the code block is labeled "Same IDs", indicating that the list object at index 1 of the tuple has the same memory address both before and after the `append` operation.

At the bottom right of the slide, there is a small logo and the URL avichawla.substack.com.

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')
```

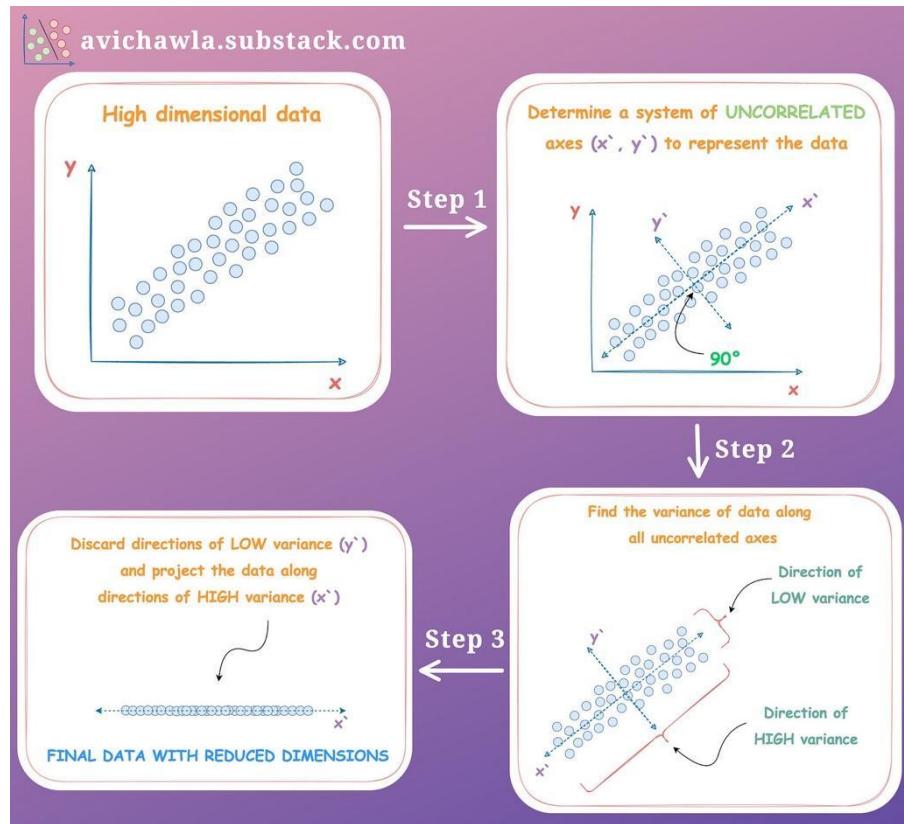
Alternatively, you can also use the following code:

```
%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](#).



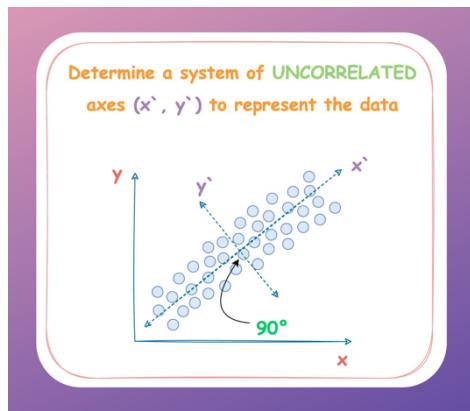
A Visual and Overly Simplified Guide to PCA



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.

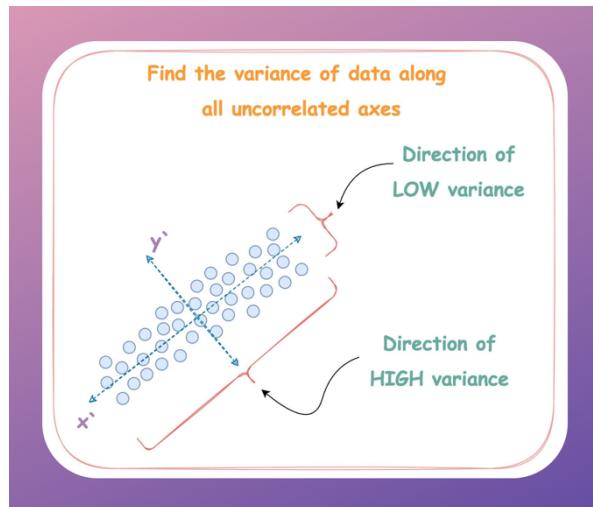
In a gist, while reducing the dimensions, the aim is to retain as much variation in data as possible.

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.

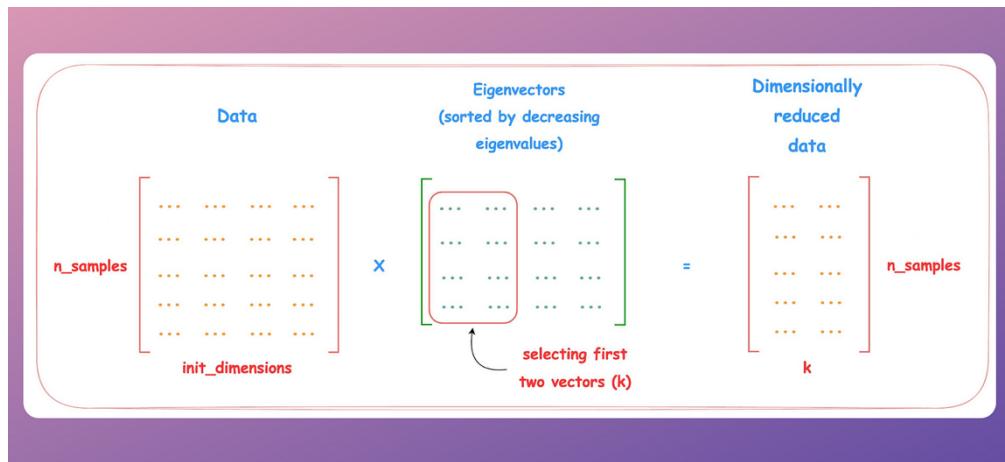
Next, we find the variance of our data along these uncorrelated axes. The variance is represented by the corresponding eigenvalues.



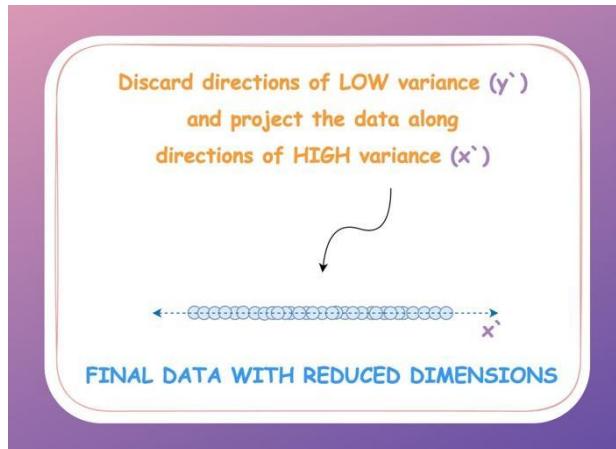
Next, we decide the number of dimensions we want our data to have post-reduction (a hyperparameter), say two. As our aim is to retain as much variance as possible, we select two eigenvectors with the highest eigenvalues.

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 [ ]:
```

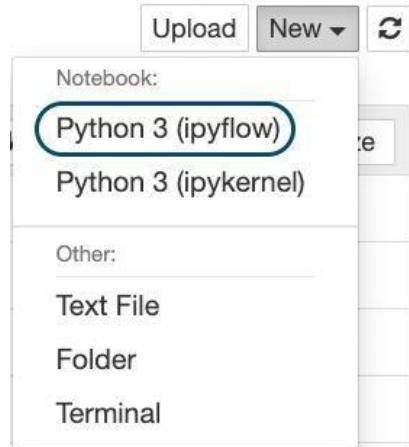
Thus, at any point, you can obtain the corresponding code to reconstruct any symbol.

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:



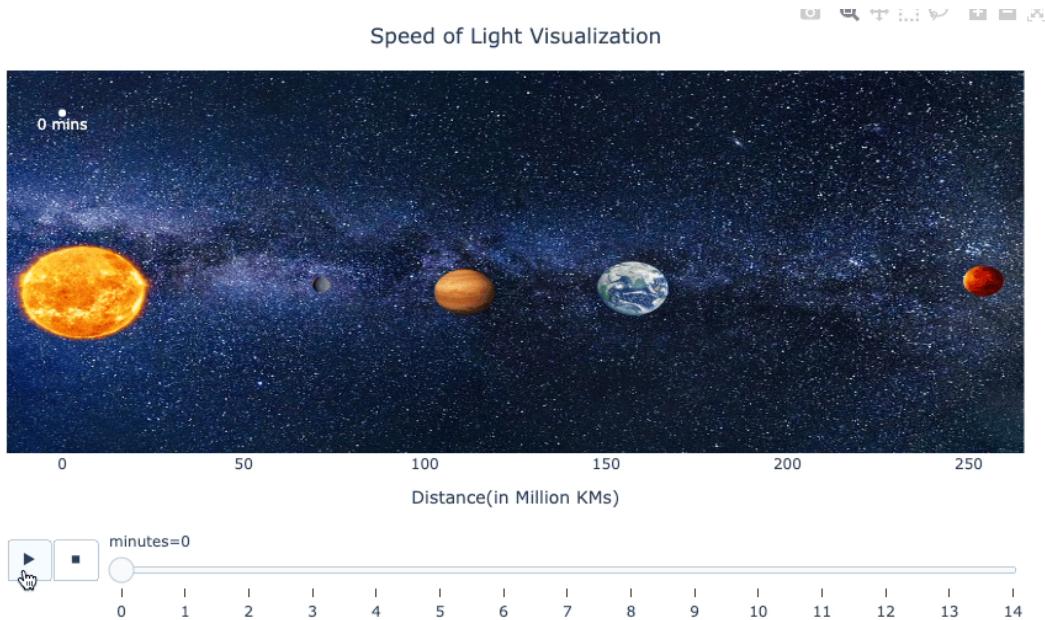
Find more details here: [ipyflow](#).



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

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")
```

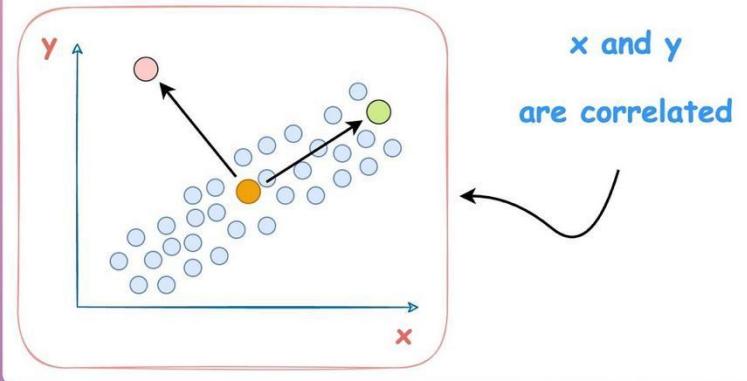
In the above function call, the data corresponding to **frame_id=0** will be plotted first. This will be replaced by the data with **frame_id=1** in the next frame, and so on.

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



The Limitation Of Euclidean Distance Which Many Often Ignore

 avichawla.substack.com



x and y are correlated

	Euclidean Distance	Mahalanobis Distance
Orange point to Green point	4.5	1.8
Orange point to Pink point	4.5	8.0
	Equal	Unequal

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.

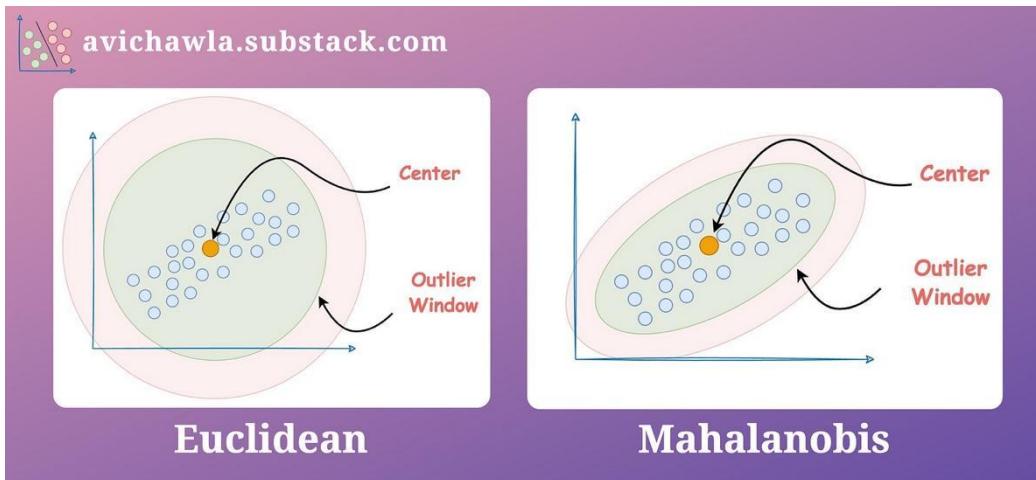
Mahalanobis distance is an excellent alternative in such cases. It is a multivariate distance metric that takes into account the data distribution.

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: $n_samples \times n_dimensions$).
- μ : mean of individual dimensions (Shape: $1 \times n_dimensions$).
- C^{-1} : Inverse of the covariance matrix
(Shape: $n_dimensions \times n_dimensions$).

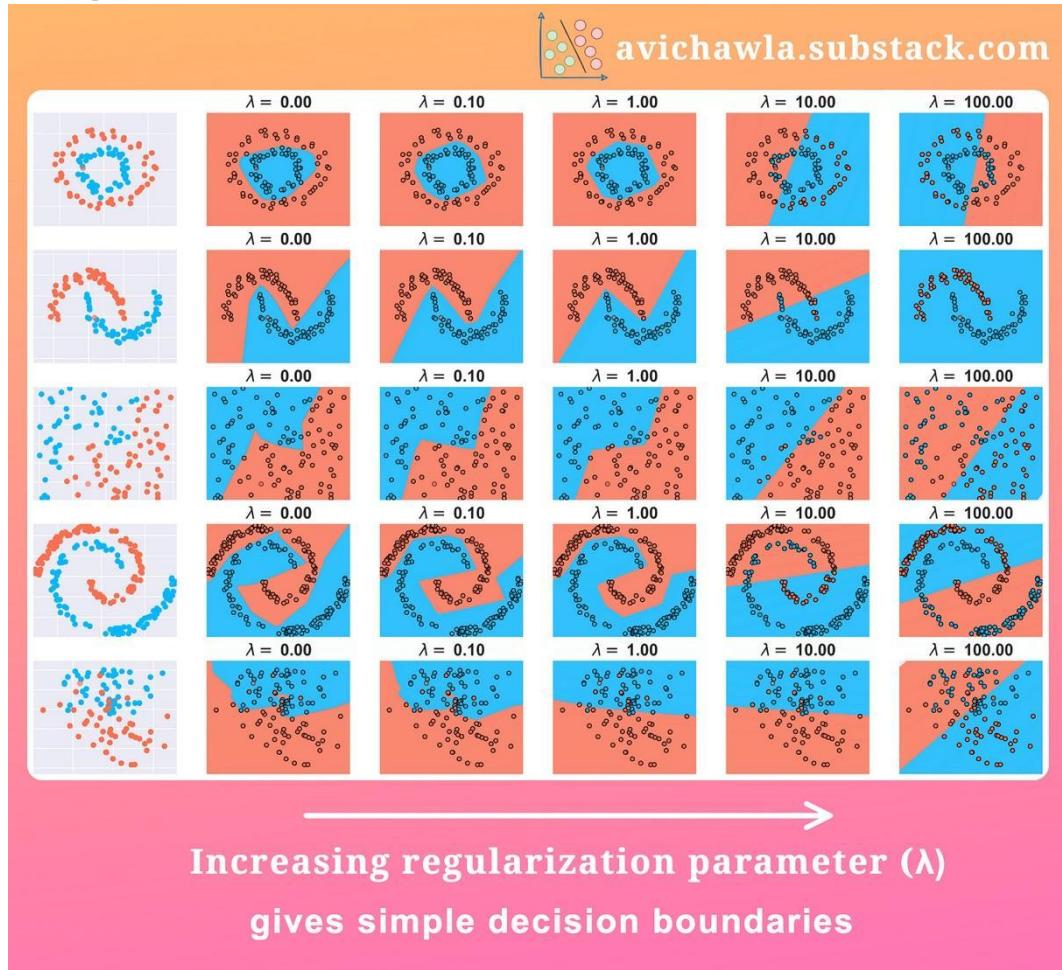


- D^2 : Square of the Mahalanobis distance
(Shape: n_samples*n_samples).

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 λ .



Cost Function = Loss + L2 Weight Penalty

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

**Higher the value of λ ,
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).

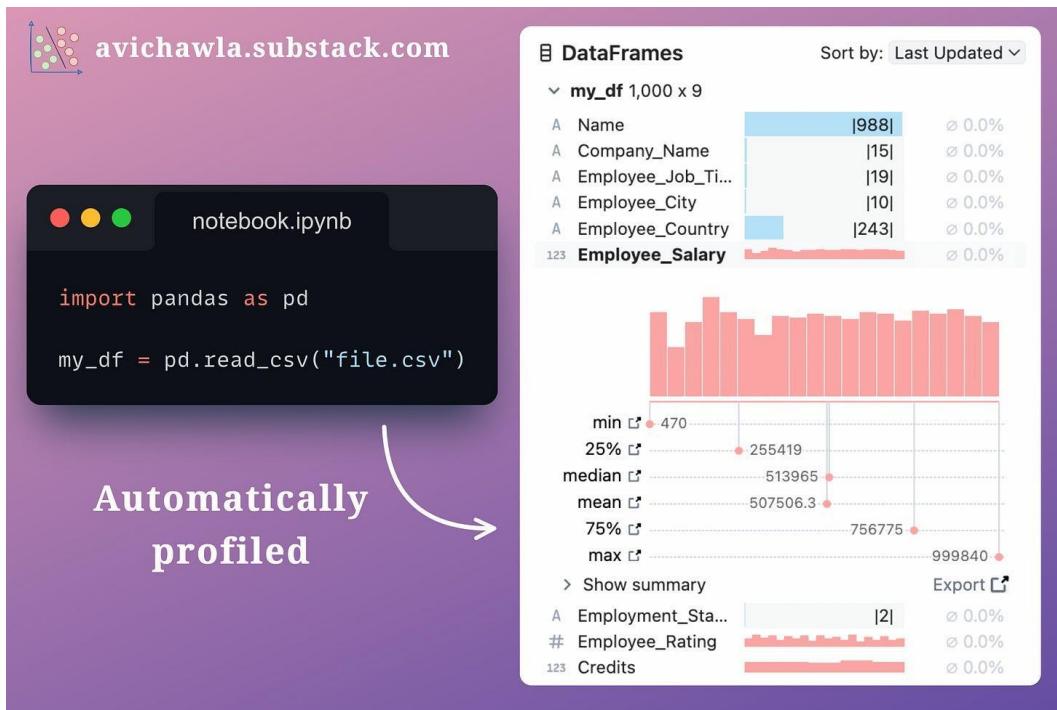
All weights close to zero

```
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	Employee_Salary	123	0.0%
Employment_Sta...	A	2	0.0%	
# Employee_Rating	123	Credits	123	0.0%

> my_df 1,000 x 9

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	1	2	(0.20%)
Renee Davis	1	2	(0.20%)
Patty Jones	1	2	(0.20%)
William Jones	1	2	(0.20%)
Daniel Lee	1	2	(0.20%)
Sarah Smith	1	2	(0.20%)
Anna Thomas	1	2	(0.20%)
Jennifer Gonzales	1	2	(0.20%)
Nicole Garcia	1	2	(0.20%)
Derek Perez	1	2	(0.20%)

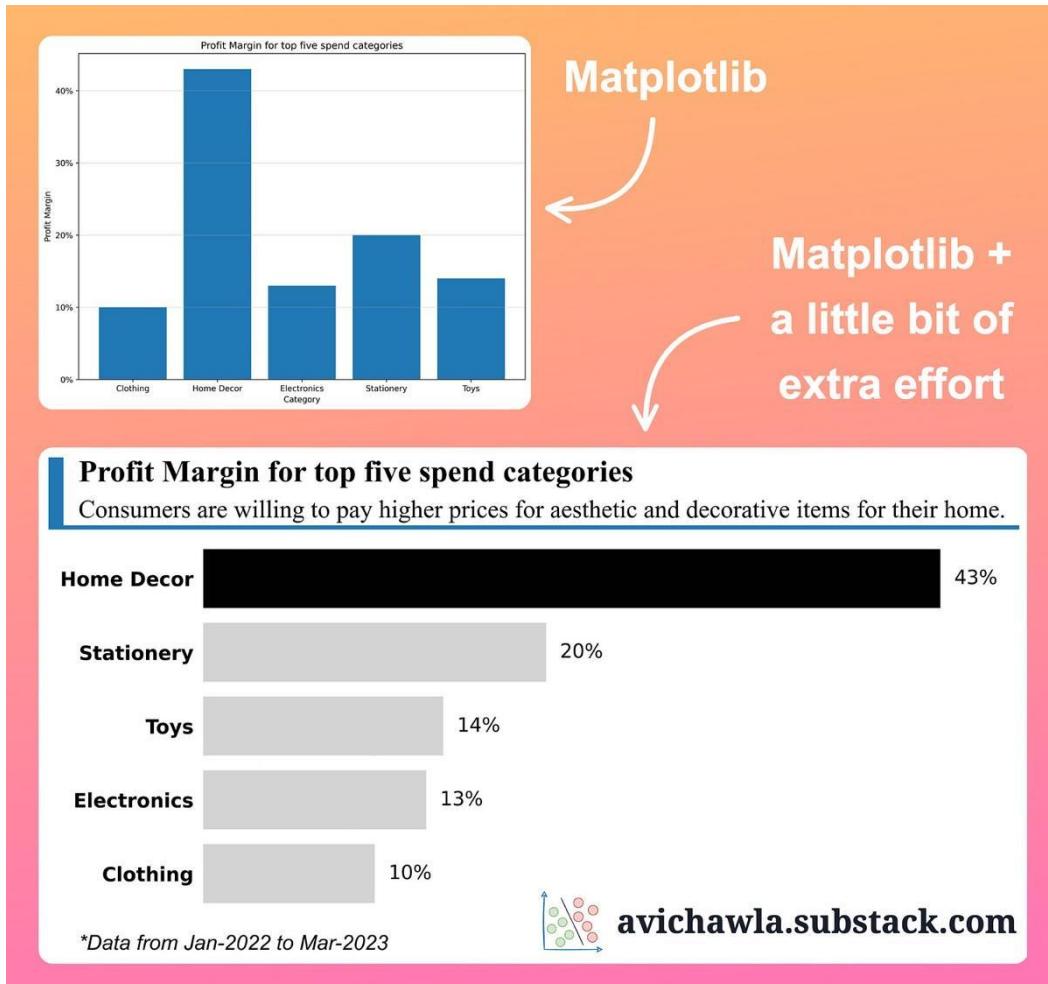
> Show summary

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	Employee_Salary	123	0.0%
Employment_Sta...	A	2	0.0%	
# Employee_Rating	123	Credits	123	0.0%

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?



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

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

```
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')
```

Annotations highlight the mutable default parameter and the resulting shared state:

- An arrow points from the text "Mutable Default Parameter" to the line `subjects = []`.
- An annotation "Appended to the same list" with three arrows points to the three occurrences of the list `['Maths']` in the output.

Output:

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

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.

```
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'],)
```

Modified default parameter



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:

The terminal window title is "Replace mutable parameter". The code defines a function `add_subject` that takes `name`, `subject`, and an optional `subjects` parameter set to `None`. It checks if `subjects` is `None` and creates an empty list if it is. Then it appends the `subject` to the list and returns a dictionary with `name` and `subjects`. The terminal shows three calls to `add_subject` with 'Maths' as the subject, resulting in three dictionaries where each has a unique list of subjects.

```
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:

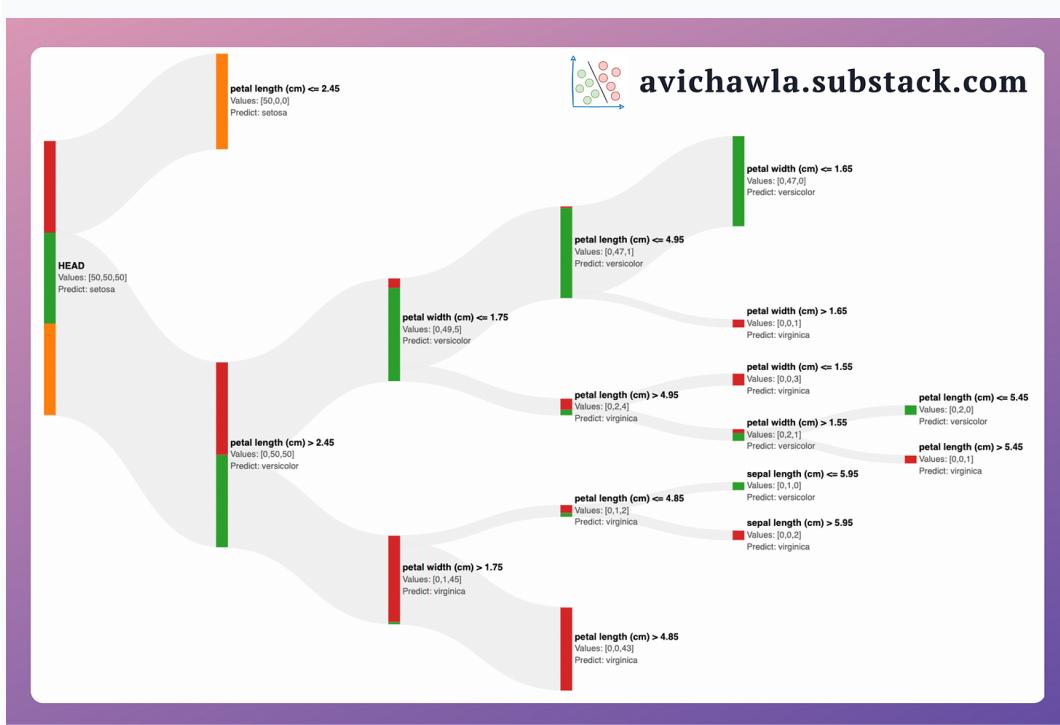
```
{'name': 'Joe', 'subjects': ['Maths']}
{'name': 'Bob', 'subjects': ['Maths']}
{'name': 'Roy', 'subjects': ['Maths']}
```

avichawla.substack.com

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



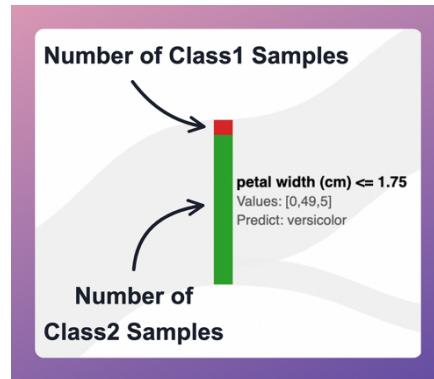
In one of my earlier posts, I explained why sklearn's decision trees always overfit the data with its default parameters (read [here](#) if you wish to recall).

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.

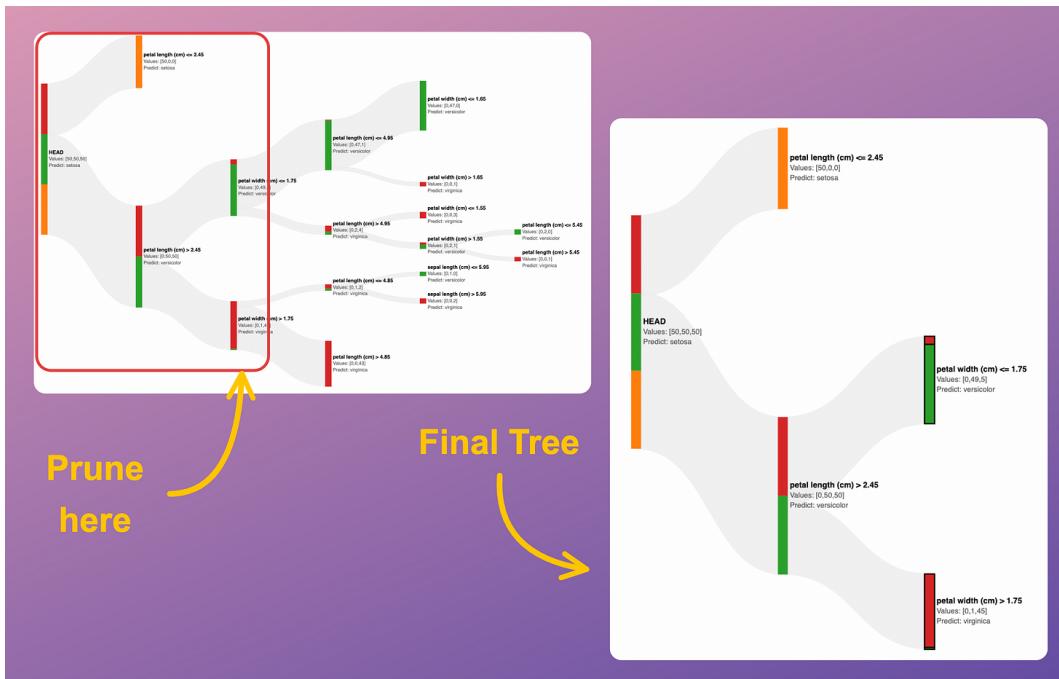
The Sankey diagram above allows you to interactively visualize the predictions of a decision tree at each node.

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.

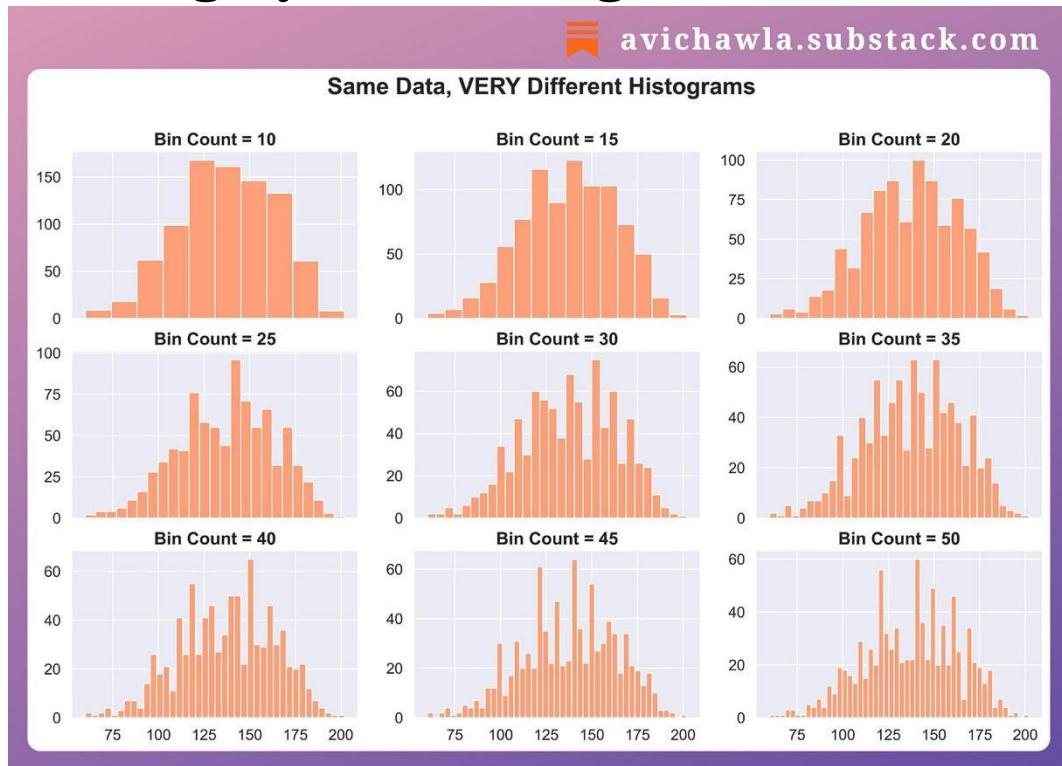
For instance, in the full decision tree shown below, pruning the tree at a depth of two appears to be reasonable.



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.

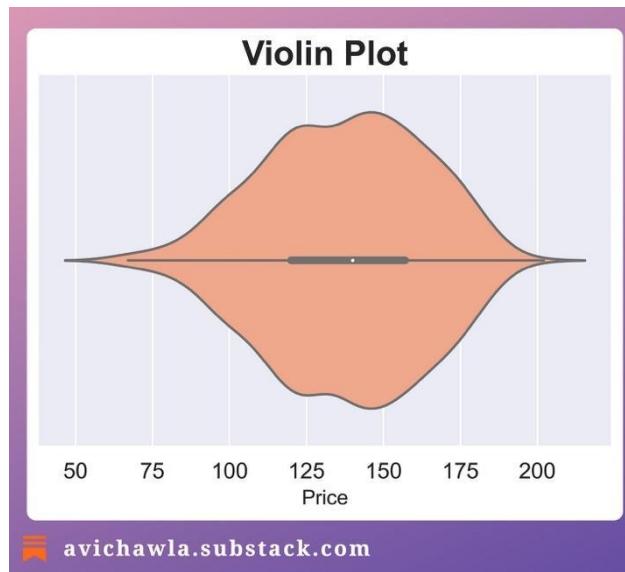
The figure above depicts the histograms obtained on the same data, but by altering the number of bins. Each histogram conveys a different story, even though the underlying data is the same.

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:



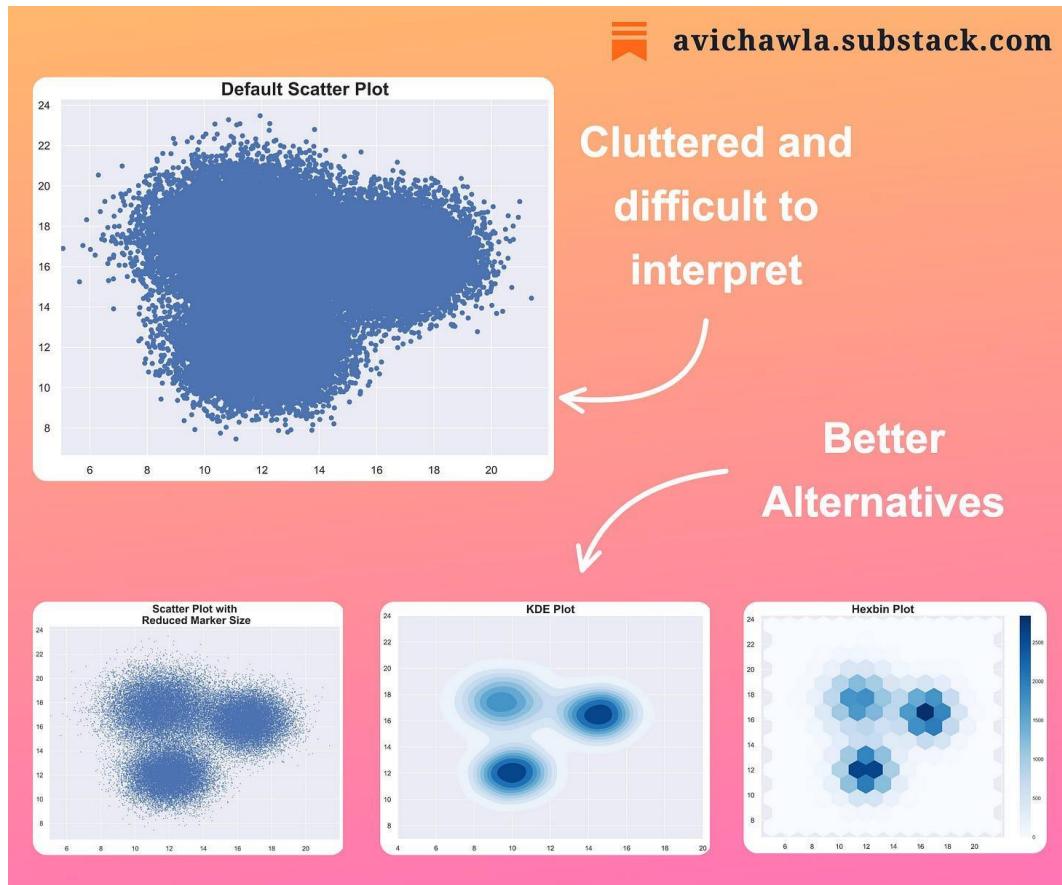
avichawla.substack.com

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.



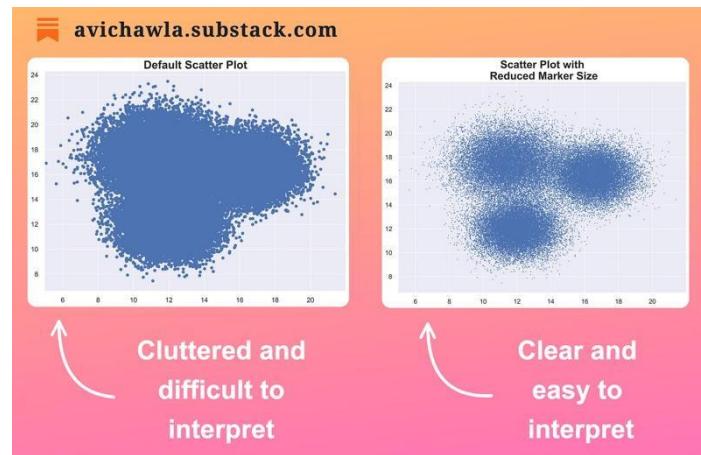
Three Simple Ways To (Instantly) Make Your Scatter Plots Clutter Free



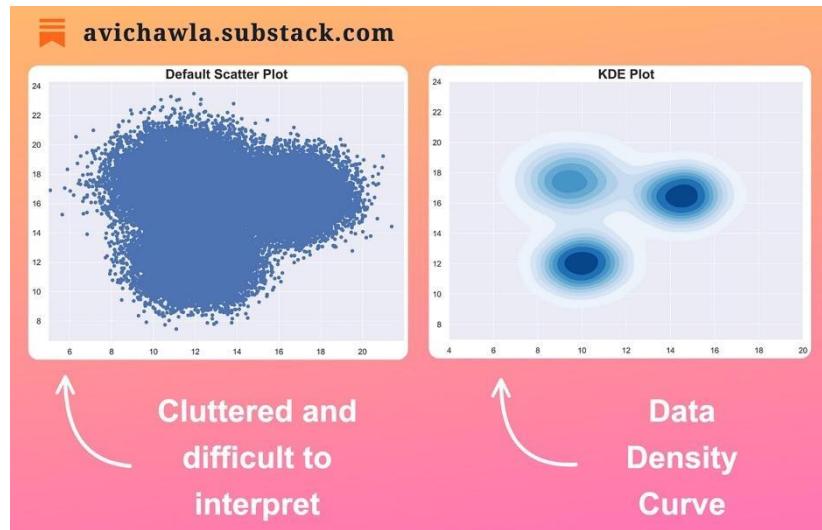
Scatter plots are commonly used in data visualization tasks. But when you have many data points, they often get too dense to interpret.

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

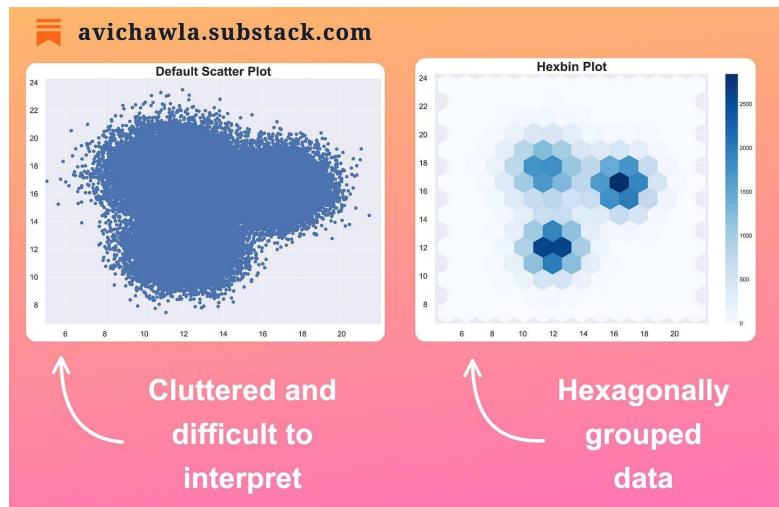
One of the simplest yet effective ways could be to reduce the marker size. This, at times, can instantly offer better clarity over the default plot.



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.

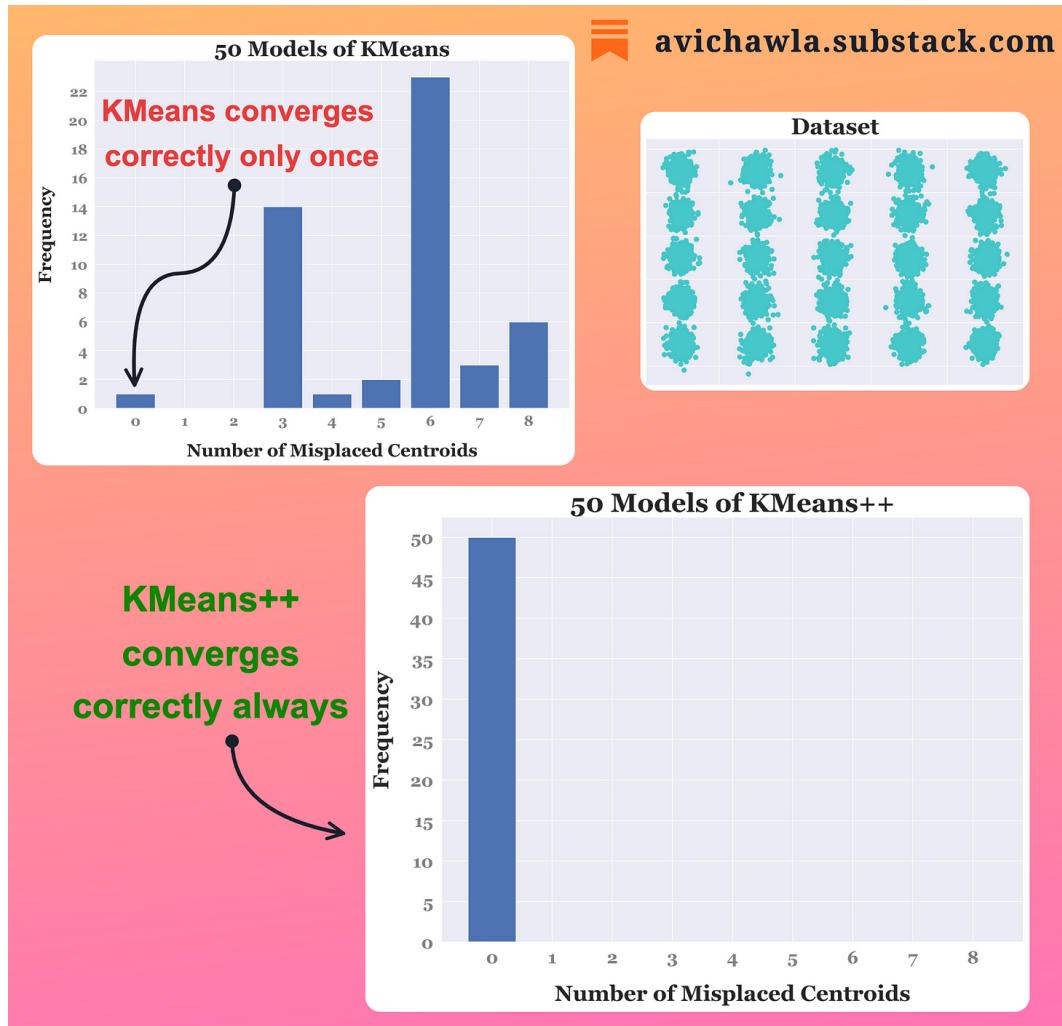


Lastly, another better alternative can be a hexbin plot. It bins the chart into hexagonal regions and assigns a color intensity based on the number of points in that area.





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



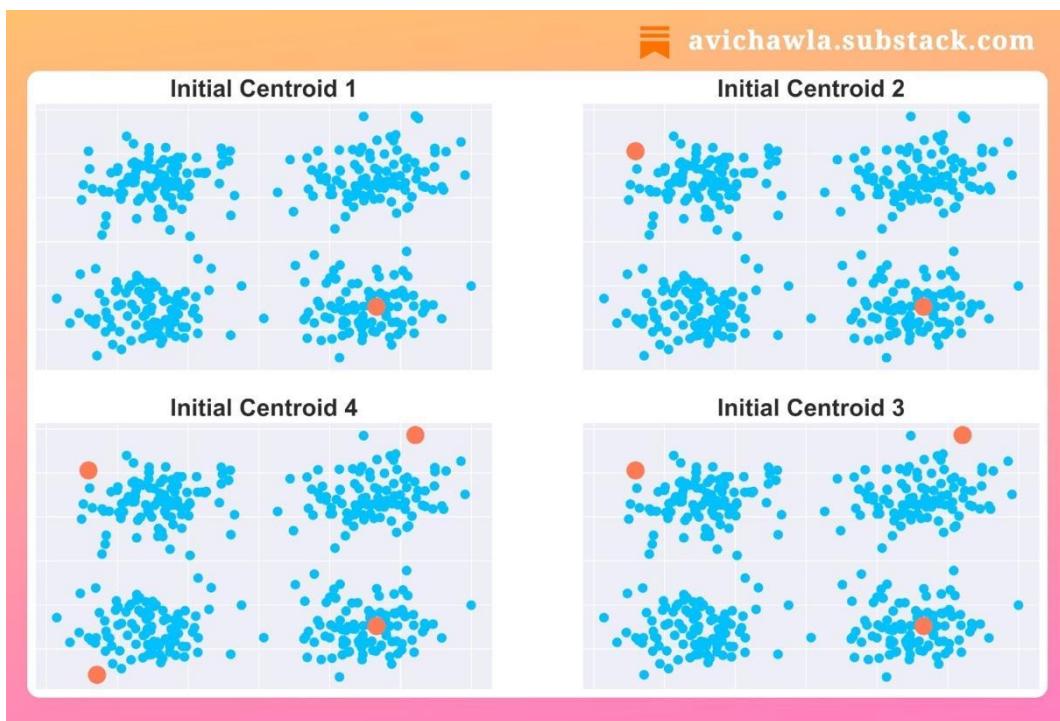
The most important yet often overlooked step of KMeans is its centroid initialization. Here's something to consider before you use it next time.

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.

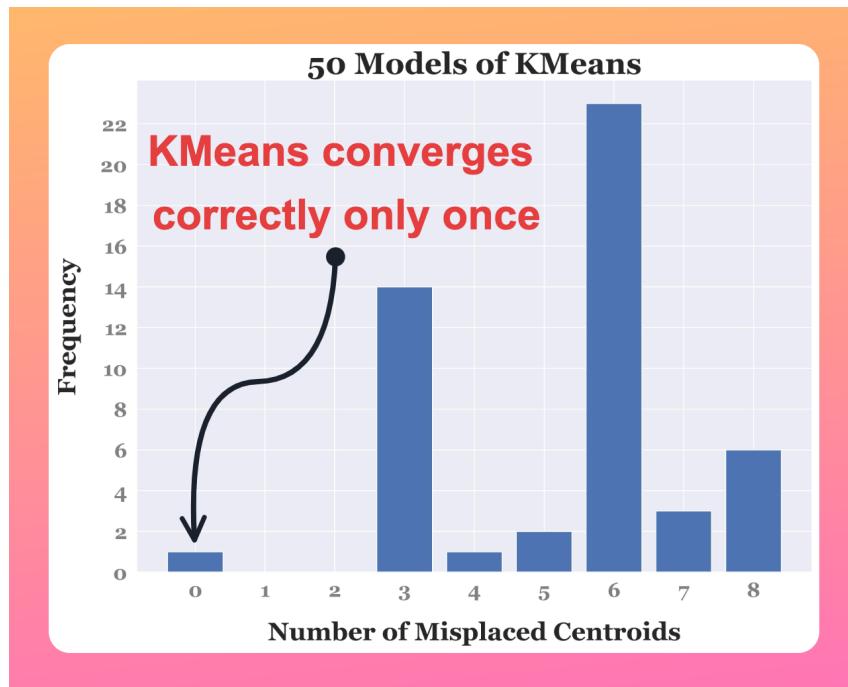
The first centroid is selected randomly. But the next centroid is chosen based on the distance from the first centroid.



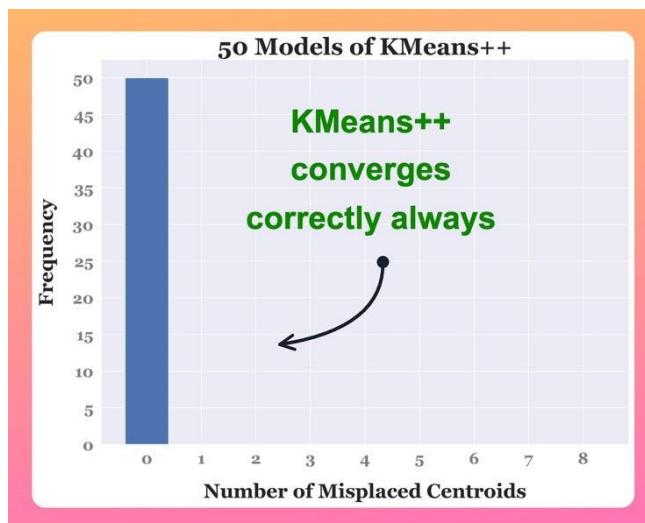
In other words, a point that is away from the first centroid is more likely to be selected as an initial centroid. This way, all the initial centroids are likely to lie in different clusters already, and the algorithm may converge faster and more accurately.

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.

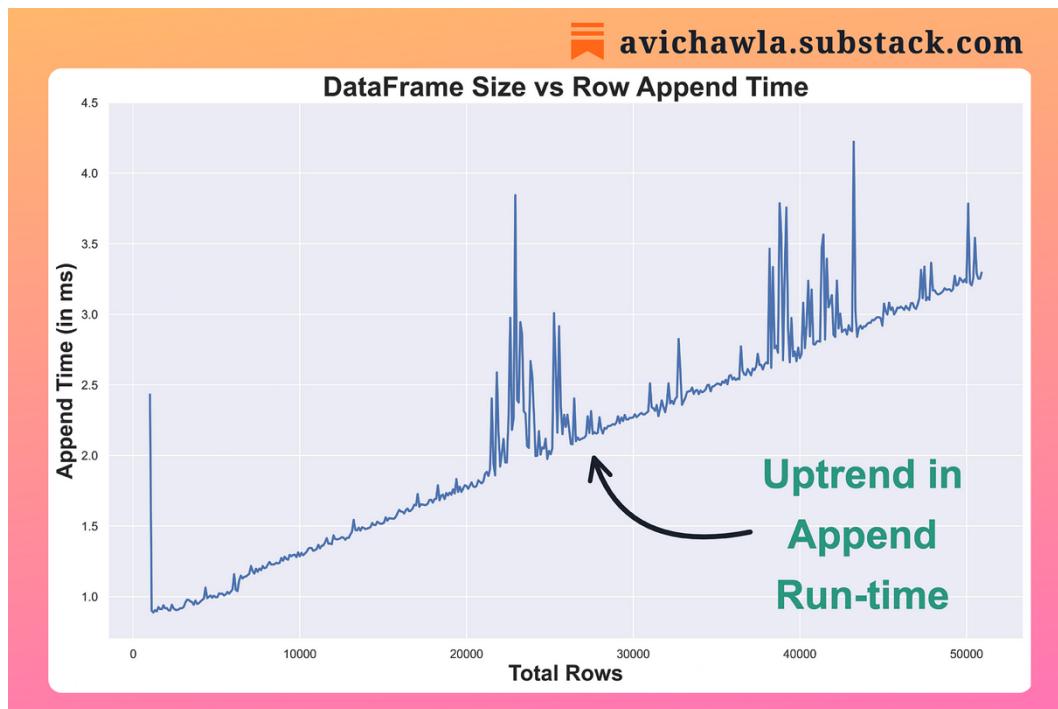


Luckily, if you are using sklearn, you don't need to worry about the initialization step. This is because sklearn, by default, resorts to the KMeans++ approach.

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

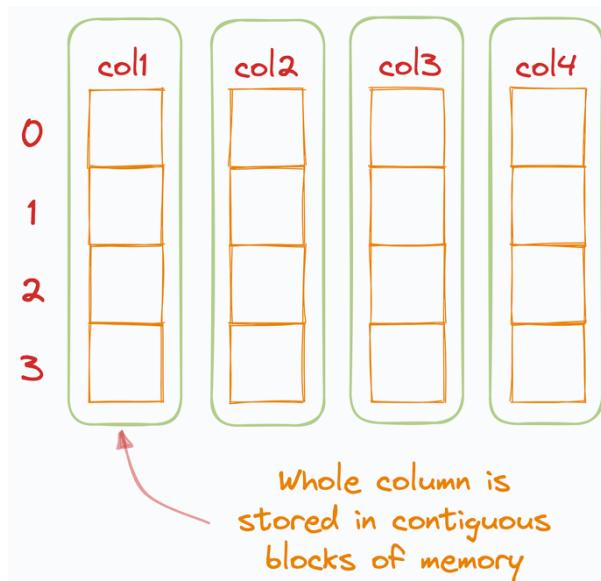


Why You Should Avoid Appending Rows To A DataFrame



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.

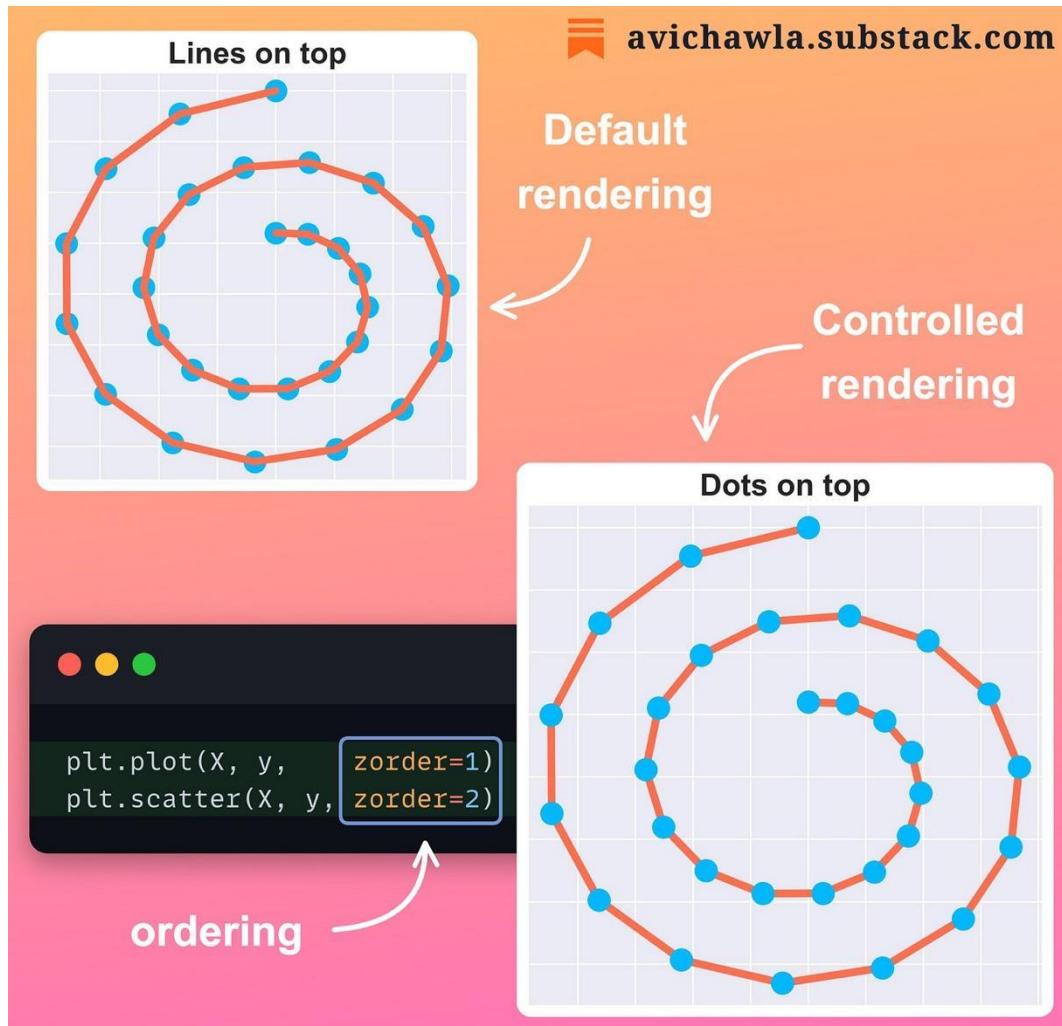
Thus, if you intend to grow a dataframe (row-wise) this frequently, it is better to first convert the dataframe to another data structure, a dictionary or a numpy array, for instance.

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.

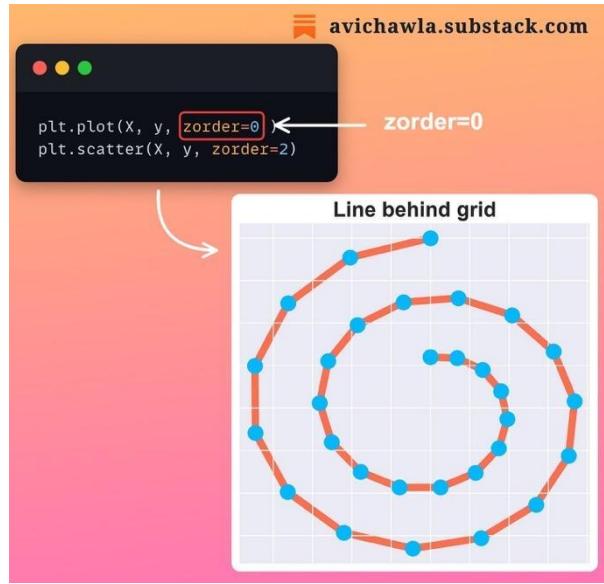
By default, matplotlib renders different types of elements (also called artists), like plots, legend, texts, etc., in a specific order.

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](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.zorder.html).



A Counterintuitive Thing About Python Dictionaries

avichawla.substack.com

```
>>> 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
```

The terminal shows two blocks of Python code. The first block defines a dictionary `my_dict` with four key-value pairs: `1.0` (float), `1` (int), `True` (bool), and `'1'` (string). The second block prints the dictionary, which only contains the key-value pairs for `1.0` and `'1'`. Handwritten-style annotations are overlaid on the right: 'Added 4 keys' points to the first block, 'dict only has 2 keys' points to the second block, and an arrow from 'dict only' to 'has' indicates the discrepancy.

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)'}
```

float key value of boolean key



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

The diagram illustrates the performance gain of using Codon over Python. It shows three windows:

- A top window titled "big_loop.py" containing Python code:

```
result = []
for a in range(10000):
    for b in range(10000):
        if (a+b)%11 == 0:
            result.append((a,b))
```
- A middle window titled "Python" showing the command and run-time:

```
$ python big_loop.py
# Run-time: 10.9s
```
- A bottom window titled "Codon" showing the command and run-time:

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

A large white arrow points from the Python run-time window to the Codon run-time window, with the text "100x Faster" written above it.

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:



avichawla.substack.com

fib.py

```
def fib(N):
    """
    Function to find the
    Nth Fibonacci number.

    fib(N) = fib(N-1) + fib(N-2)
    """
    ...
```

pi.py

```
def pi_approx(n_terms):
    """
    Function to find the
    approximate value of pi.

    pi = 4*(1 - 1/3 + 1/5 - 1/7...)
    """
    ...
```

Python

```
$ python fib.py # N=35
# Time: 2.53s
```

```
$ python fib.py # N=45
# Time: 296s
```

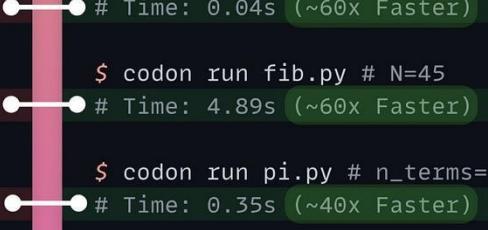
```
$ python pi.py # n_terms=10^8
# Time: 14.7s
```

Codon

```
$ codon run fib.py # N=35
# Time: 0.04s (~60x Faster)
```

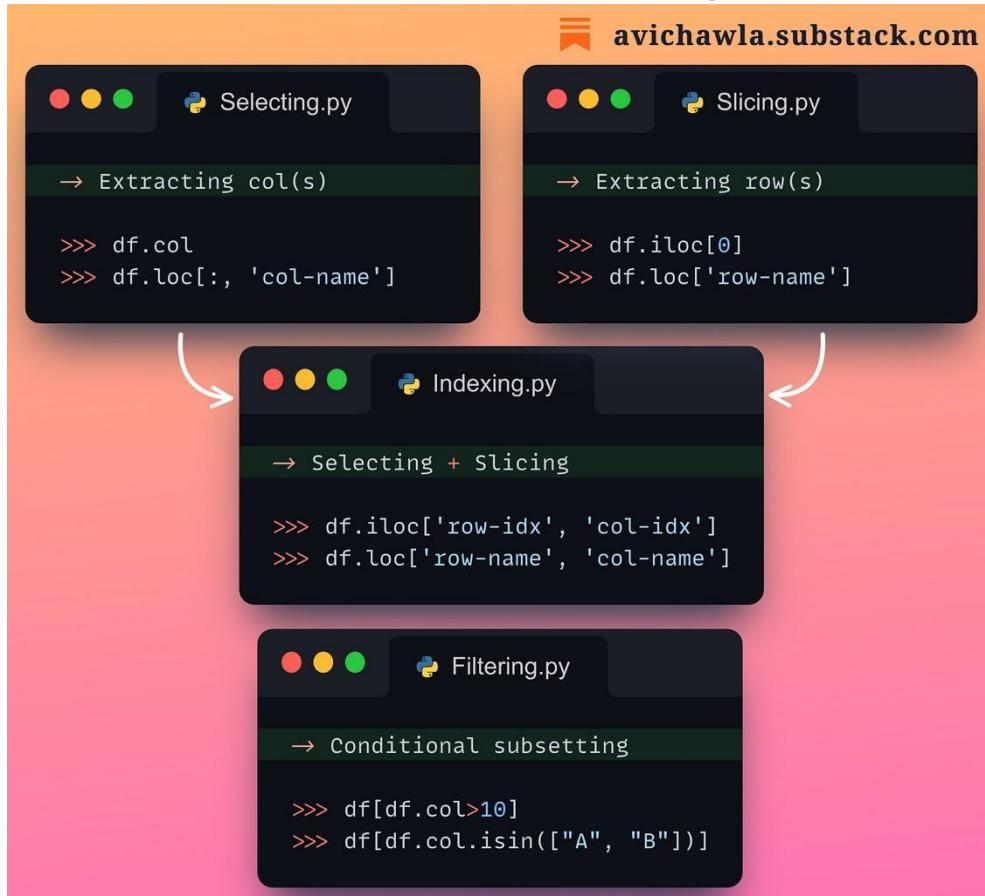
```
$ codon run fib.py # N=45
# Time: 4.89s (~60x Faster)
```

```
$ codon run pi.py # n_terms=10^8
# Time: 0.35s (~40x Faster)
```





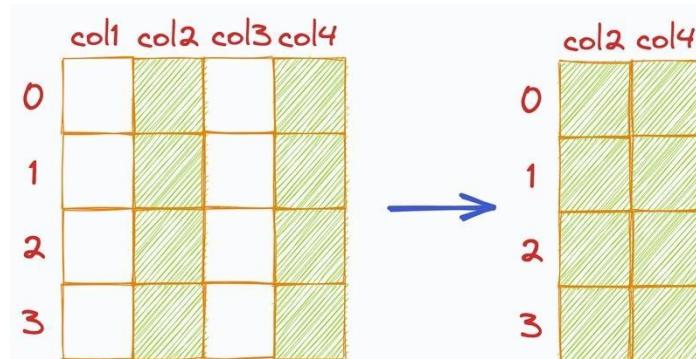
Are You Sure You Are Using The Correct Pandas Terminologies?



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

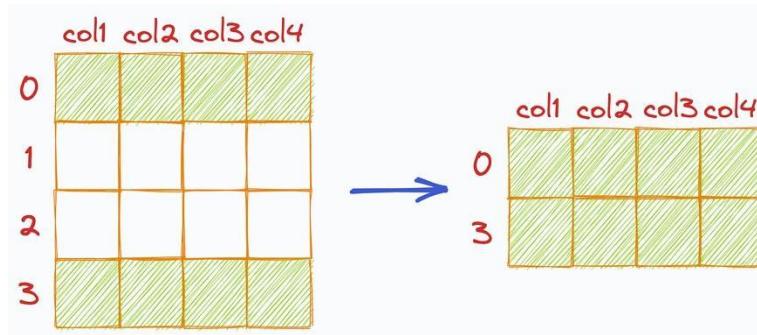
SUBSETTING means extracting value(s) from a dataframe. This can be done in four ways:

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.

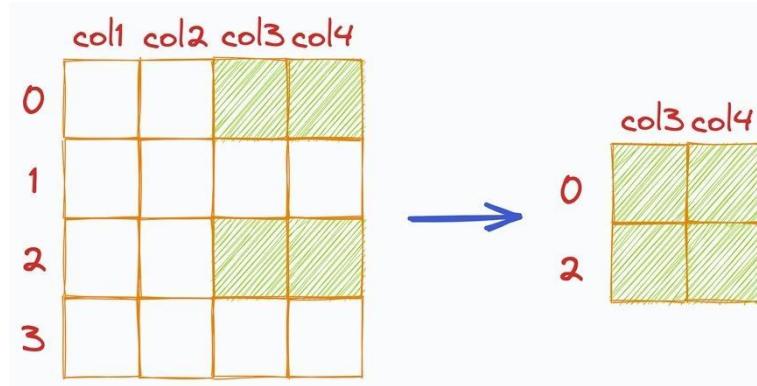




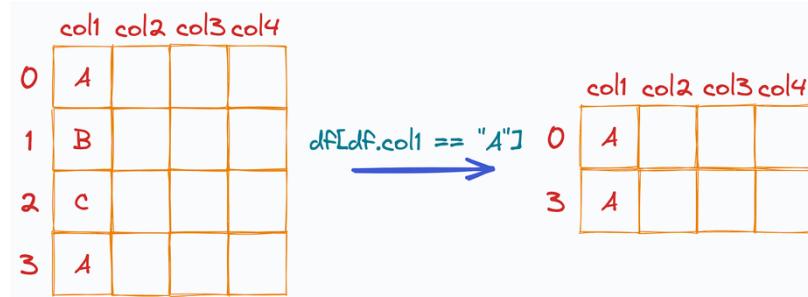
2) We call it **SLICING** when we extract one or more of its **ROWS** based on index location or name. The output contains some rows and all columns.



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?

 avichawla.substack.com

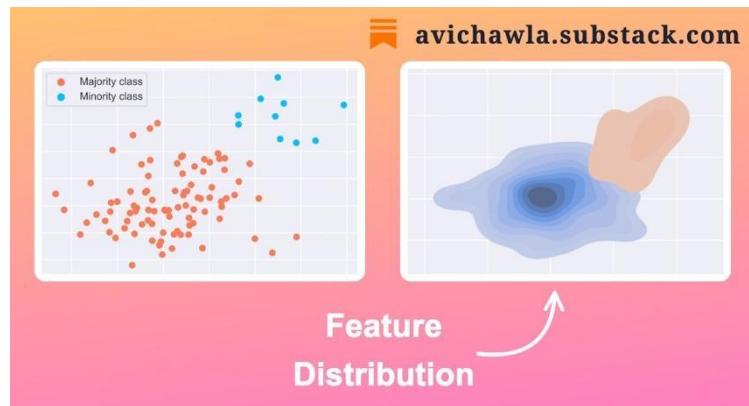


Addressing class imbalance is often a challenge in ML. Yet, it may not always cause a problem. Here's why.

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.

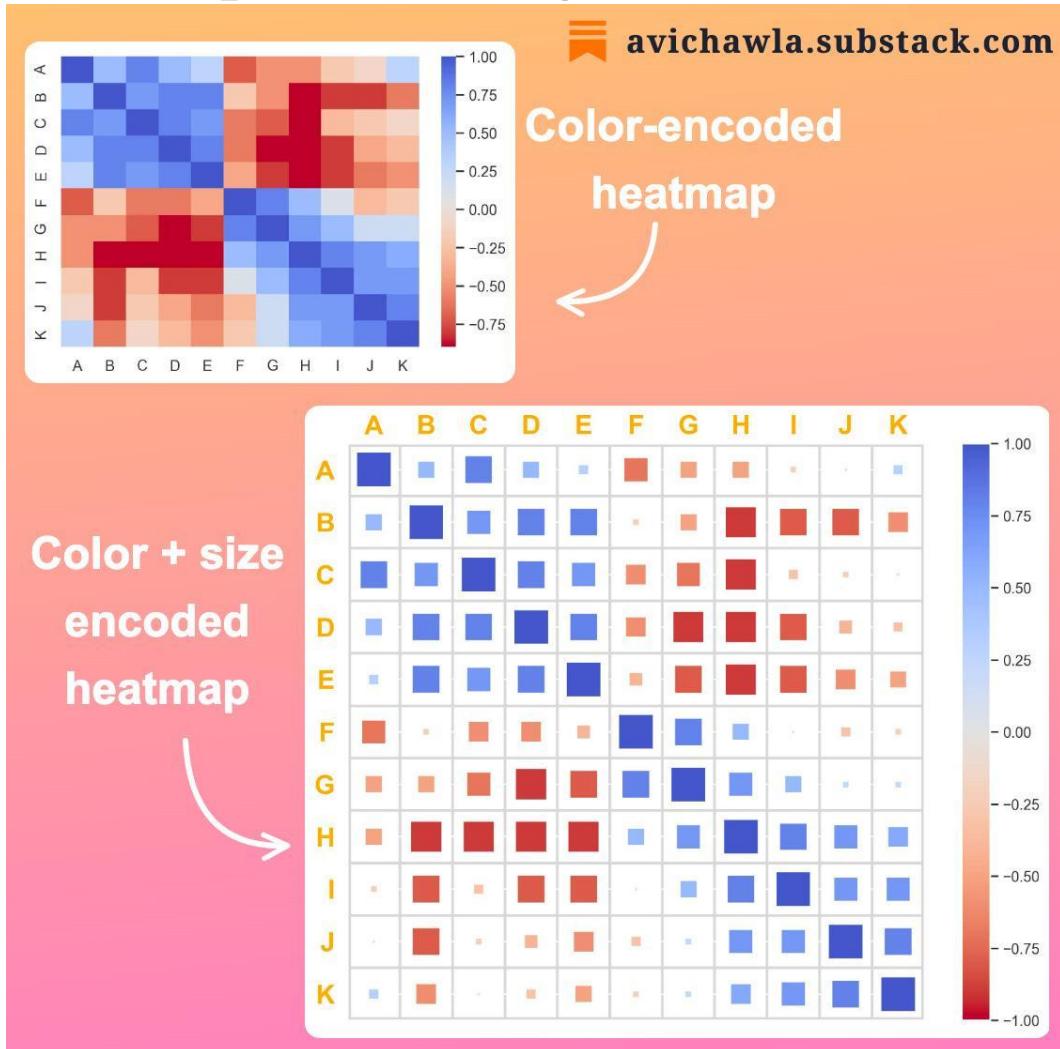
This can be done visually or by evaluating imbalance-specific metrics on simple models.

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



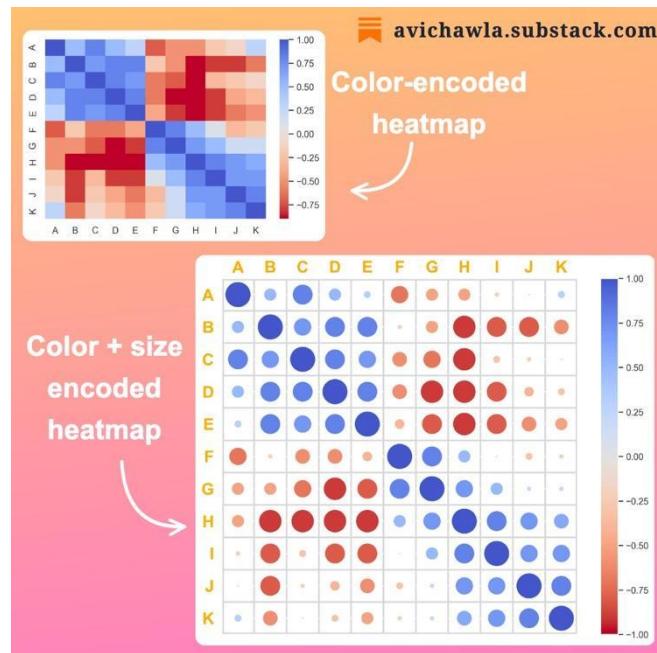
Heatmaps often make data analysis much easier. Yet, they can be further enriched with a simple modification.

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.

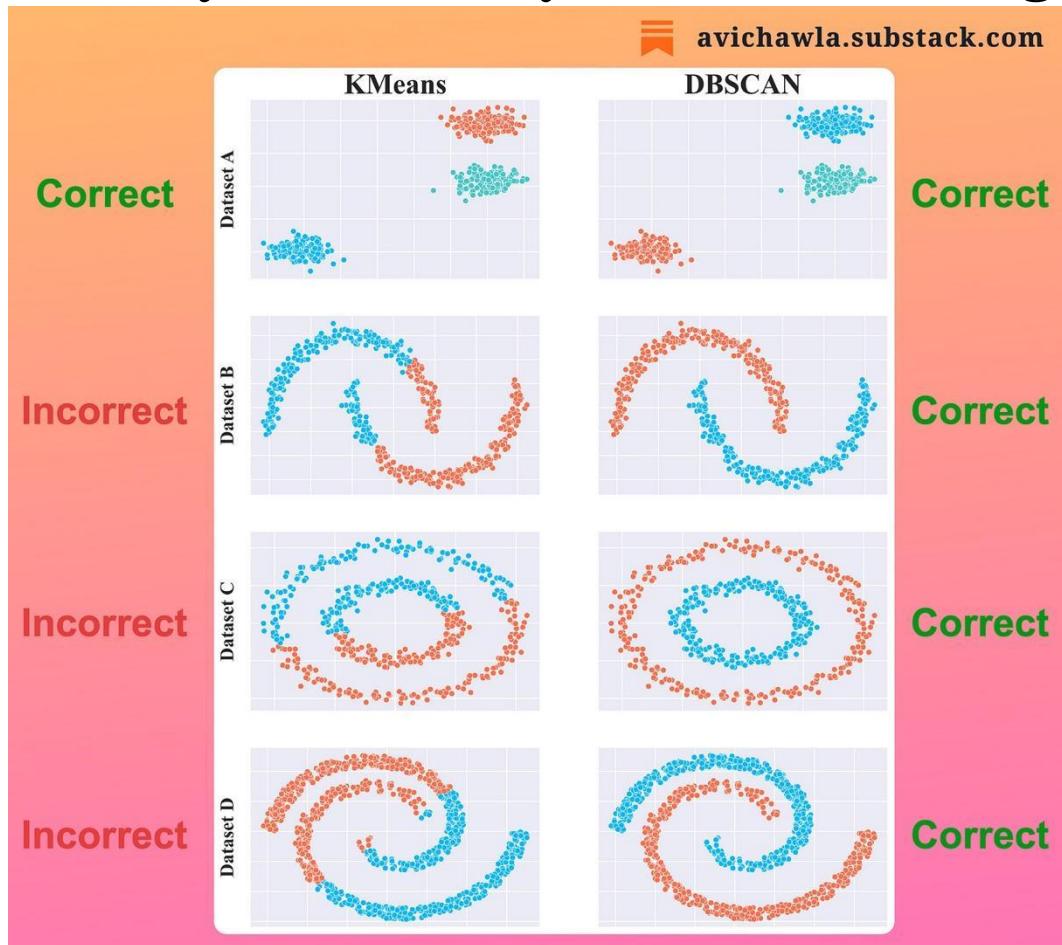
In fact, you can represent the size with any other shape. Below, I created the same heatmap using a circle instead:



Find the code for this post here: [GitHub](https://github.com/avichawla/color_size_encoded_heatmap).



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.

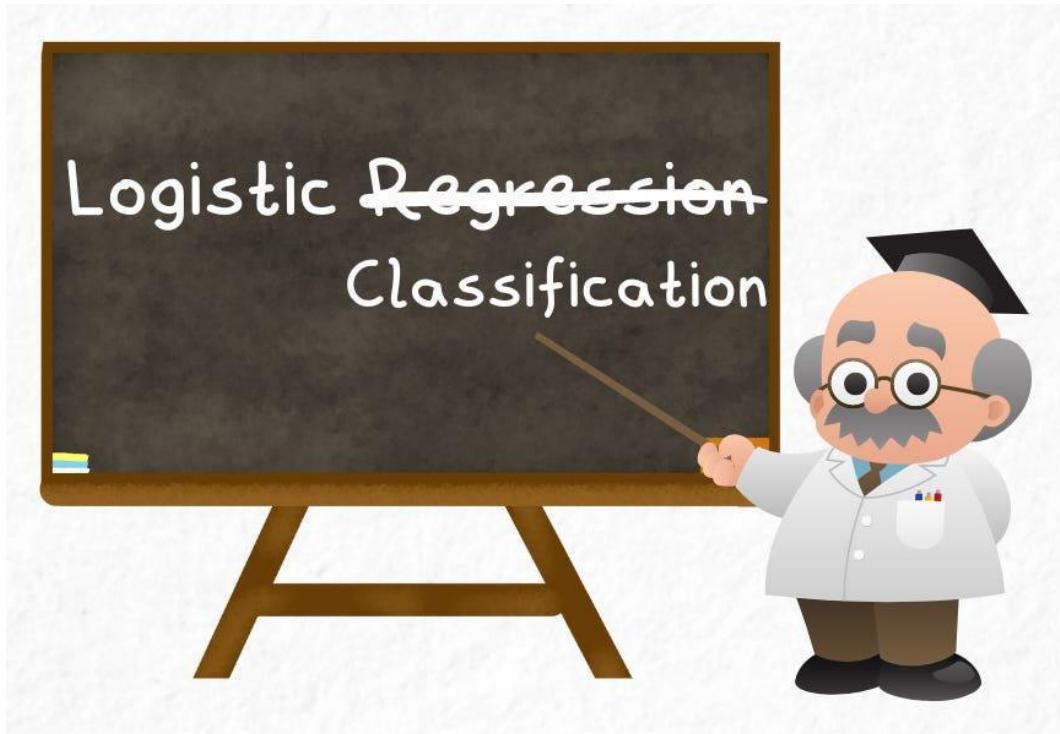
The image depicts a comparison of KMeans vs. DBSCAN on multiple datasets.

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](#).

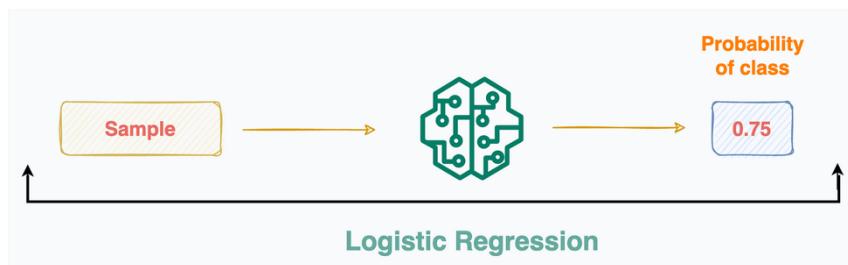


Why Don't We Call It Logistic Classification Instead?

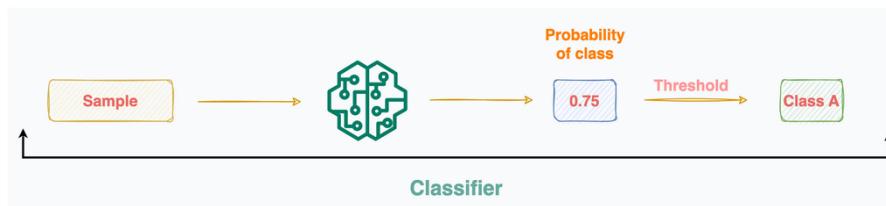


Have you ever wondered why logistic regression is called "regression" when we only use it for classification tasks? Why not call it "logistic classification" instead? Here's why.

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.



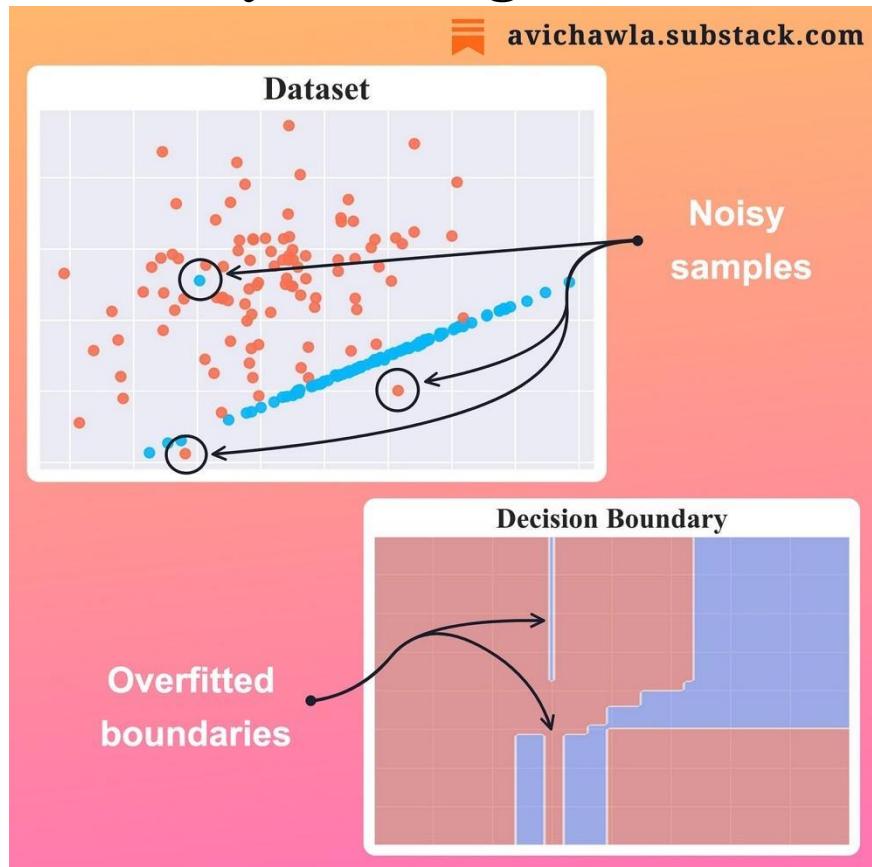
It is only when we apply those thresholds and change the interpretation of its output that the whole pipeline becomes a classifier.



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.

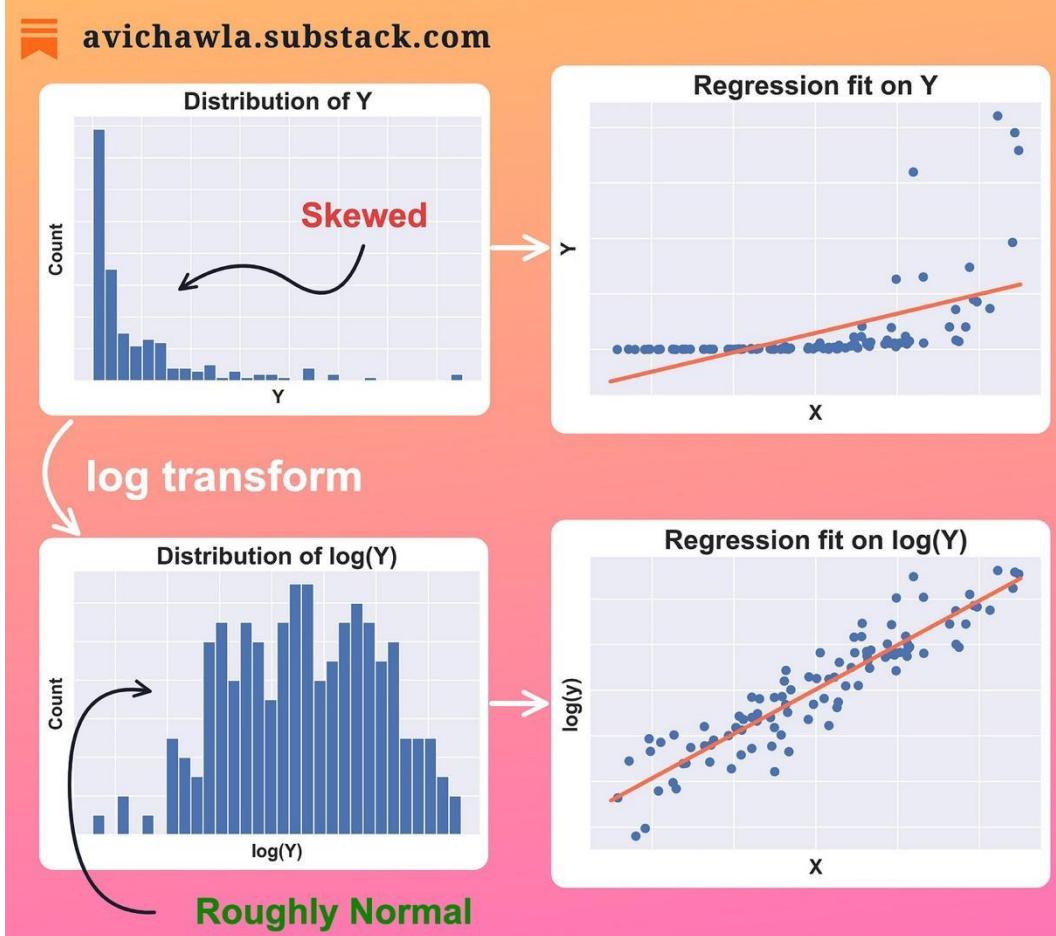
There are various techniques to avoid this, such as pruning and ensembling. Also, make sure that you tune hyperparameters if you use sklearn's implementation.

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

The screenshot shows a Substack post by [avichawla.substack.com](#). The code demonstrates various operations on a function object:

```
# 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.

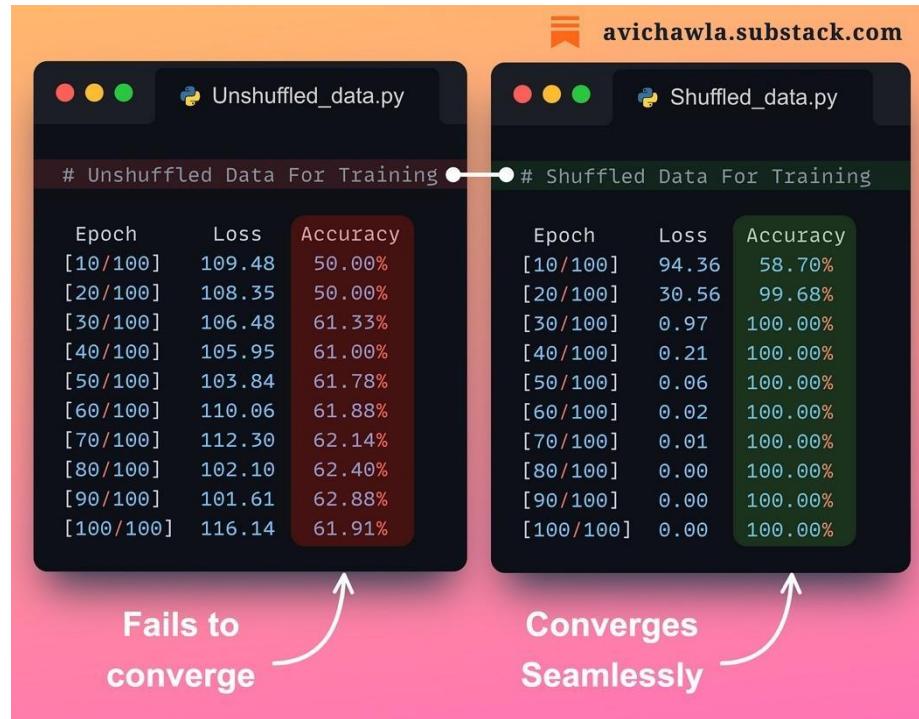
Here are a few ways to verify that python functions are indeed objects.

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.

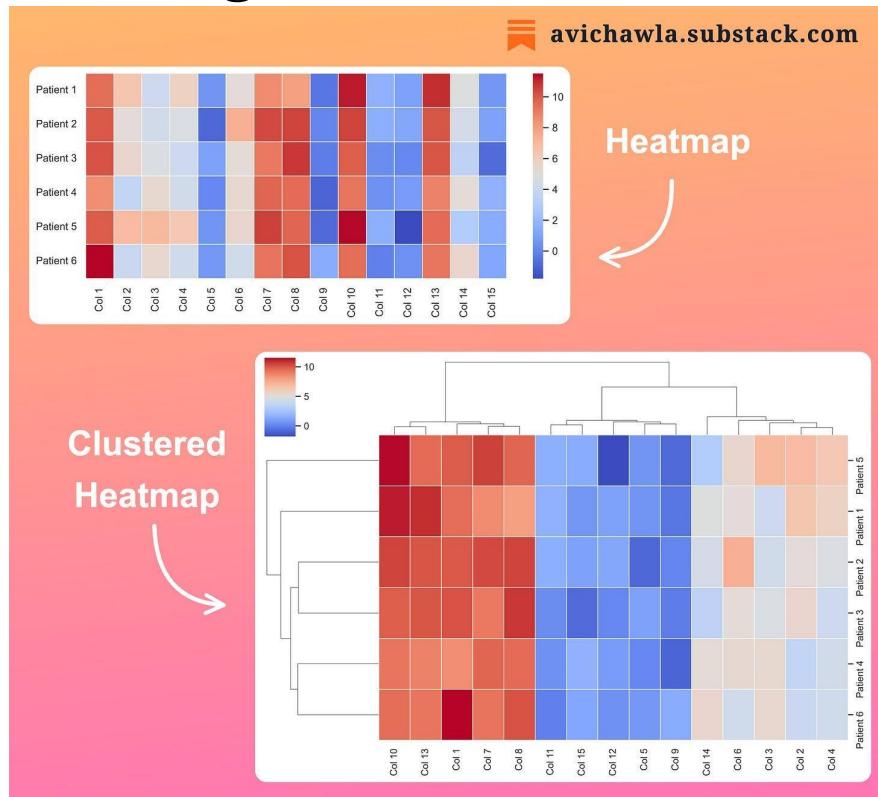
As shown, the model receiving a label-ordered dataset fails to converge. However, shuffling the dataset allows the network to learn from a more representative data sample in each batch. This leads to better generalization and performance.

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.

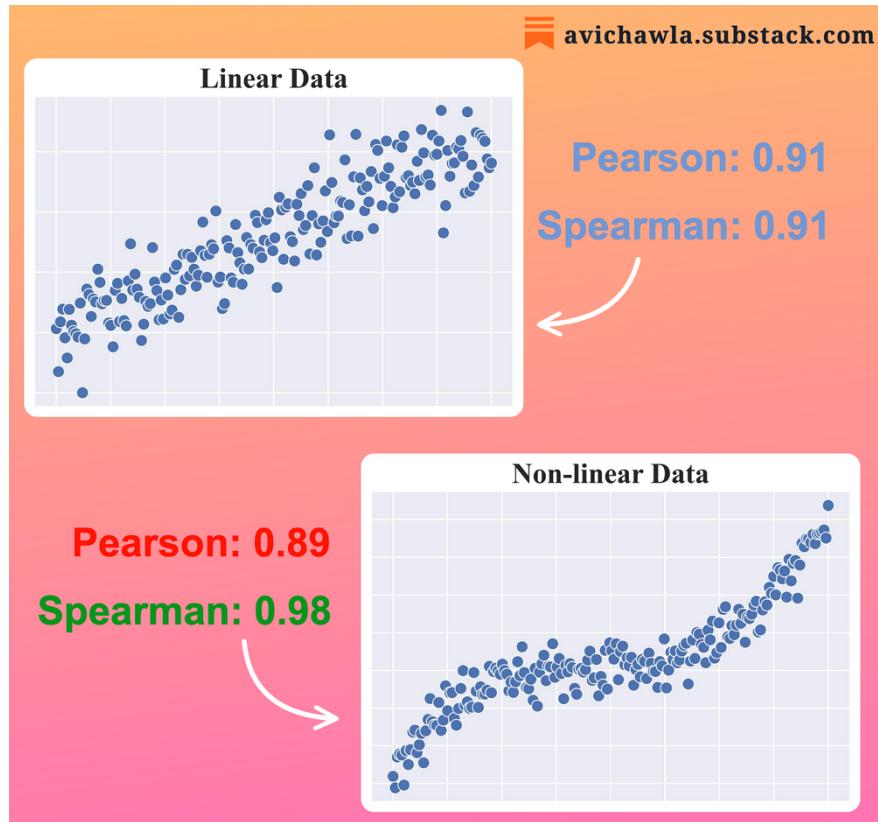
They can be especially useful when dealing with large datasets. While a traditional heatmap will be visually daunting to look at.

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.

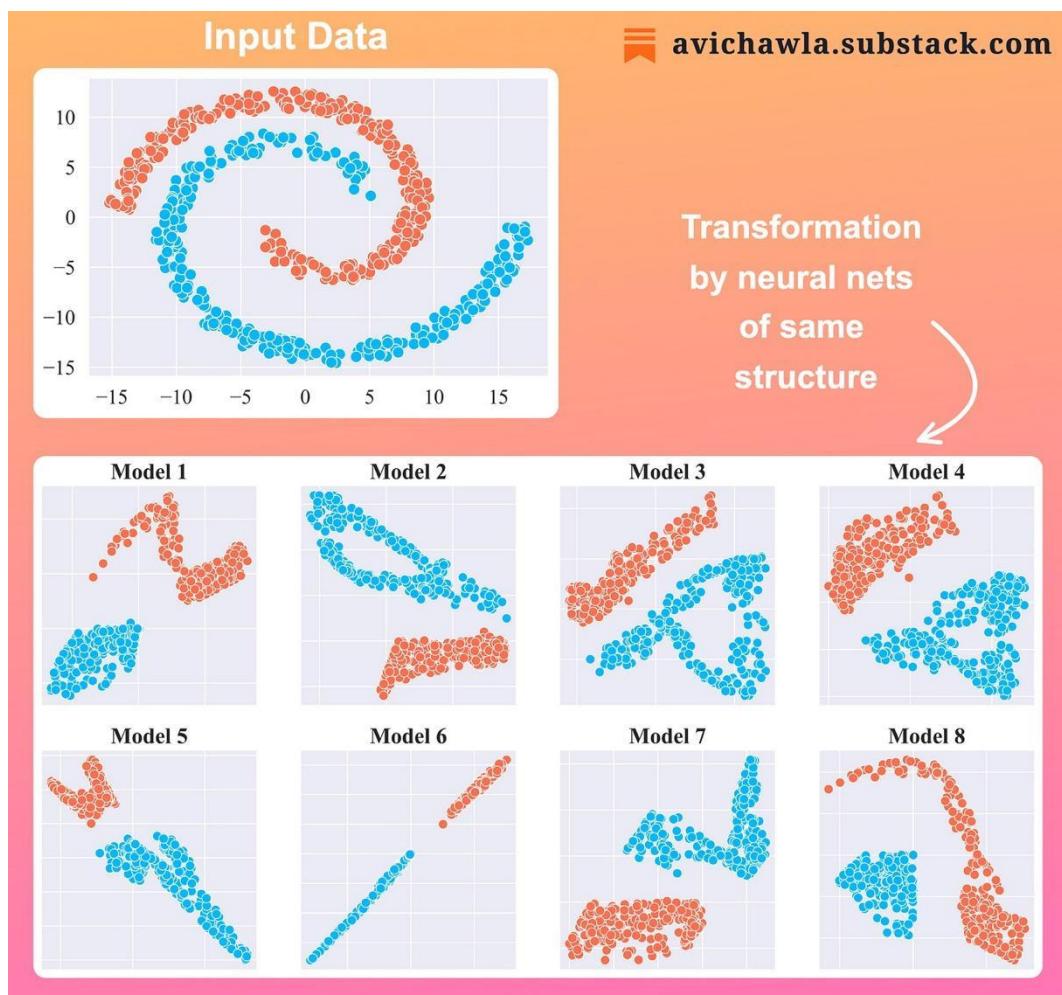
Pearson correlation primarily measures the LINEAR relationship between two variables. As a result, even if two variables have a non-linear but monotonic relationship, Pearson will penalize that.

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](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.annotate.html).



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

The screenshot shows a Jupyter notebook interface. At the top, there's a header with a bookmark icon and the URL avichawla.substack.com. Below the header, the notebook title is 'notebook.ipynb'. A code cell contains the following Python code:

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

Below the code cell is a generated visualization. It has a header '[pop out]'. On the left is a sidebar with dropdown menus for 'Table', 'Count', and 'Employment_Status'. The main area displays a pivot table with three columns: 'Employee_City', 'Employment_Status', and 'Totals'. The data is as follows:

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

(100000, 9)

1) iterrows()

```
%timeit [my_func(row) for index, row in df.iterrows()]
```

2.63 s ± 7.55 ms per loop

Slowest

2) apply()

```
%timeit df.apply(my_func, axis = 1)
```

923 ms ± 6 ms per loop

Slow

3) itertuples()

```
%timeit [my_func(row) for row in df.itertuples()]
```

87.3 ms ± 486 µs per loop

Fast

4) to_numpy()

```
%timeit np_arr = df.to_numpy(); [my_func(row) for row in np_arr]
```

32.9 ms ± 240 µs per loop

Fastest

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?

df.shape

(32768000, 9)

Access column

```
%timeit df["my_column"]
```

1.73 μ s ± 546 ns per loop

Access row

```
%timeit df.iloc[0]
```

38.4 μ s ± 1.47 μ s per loop

Column access is over 20x faster

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

avichawla.substack.com

```
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 a Substack blog post with two code snippets. The first snippet, `Method1.py`, demonstrates that modifying a list creates a new object:`1 # 1) Check ID
2 >>> id(a), id(b)
3 (12345, 12345)`

`id(a)` changed

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

The second snippet, `Method2.py`, demonstrates that modifying a list using += updates the same object:`1 # 1) Check ID
2 >>> id(a), id(b)
3 (12345, 12345)`

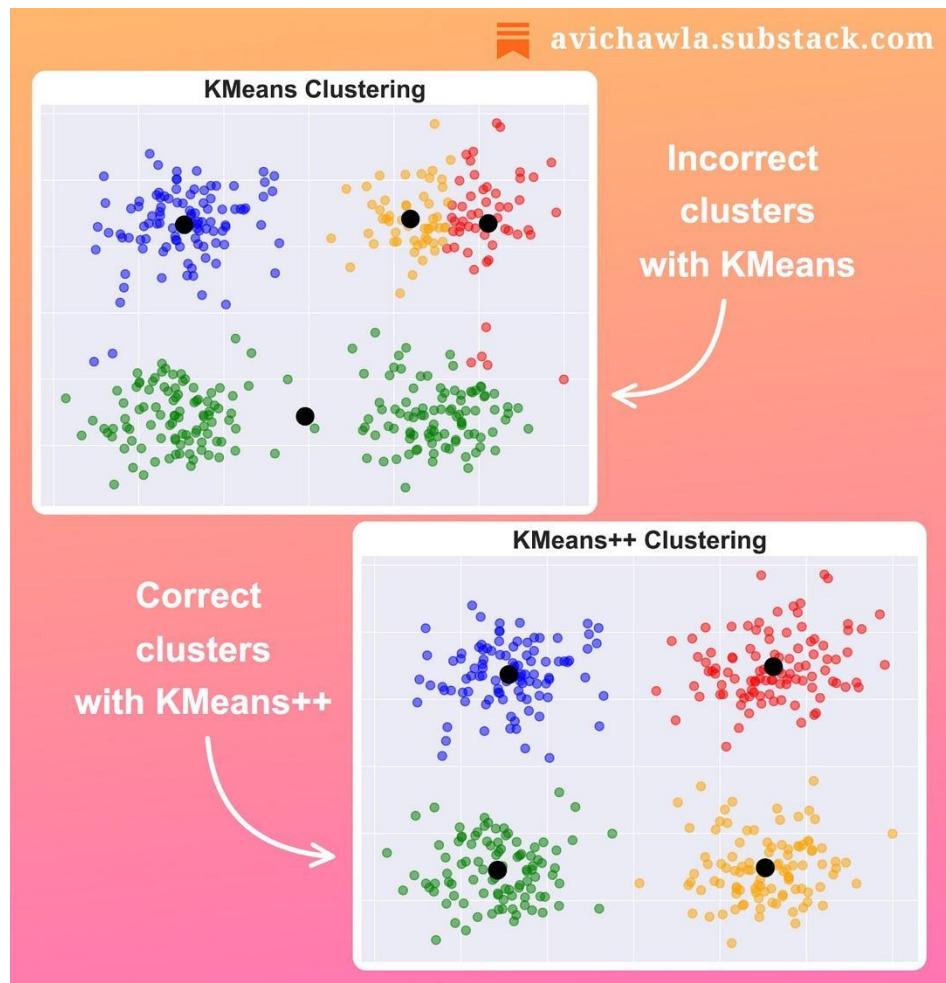
`id(a)` unchanged

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

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.

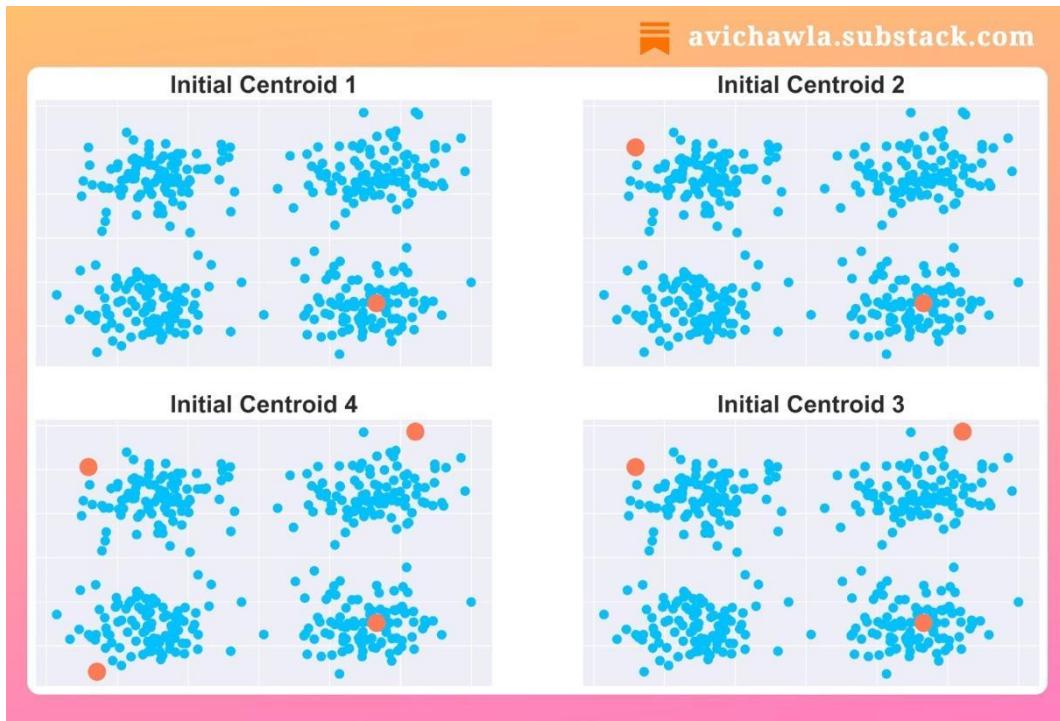
KMeans selects the initial centroids randomly. As a result, it fails to converge at times. This requires us to repeat clustering several times with different initialization.

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

The screenshot shows a Substack post by [avichawla.substack.com](#). The code defines a `Point2D` class with `__new__` and `__init__` methods. The `__new__` method checks if `x` and `y` are integers, prints "Creating Object!", and returns a new object. If not integers, it raises a `TypeError`. The `__init__` method initializes `x` and `y` and prints "Object Initialized!". When creating `p1`, both `__new__` and `__init__` logs are shown. Creating `p2` with non-integer values results in a `TypeError`.

```
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.

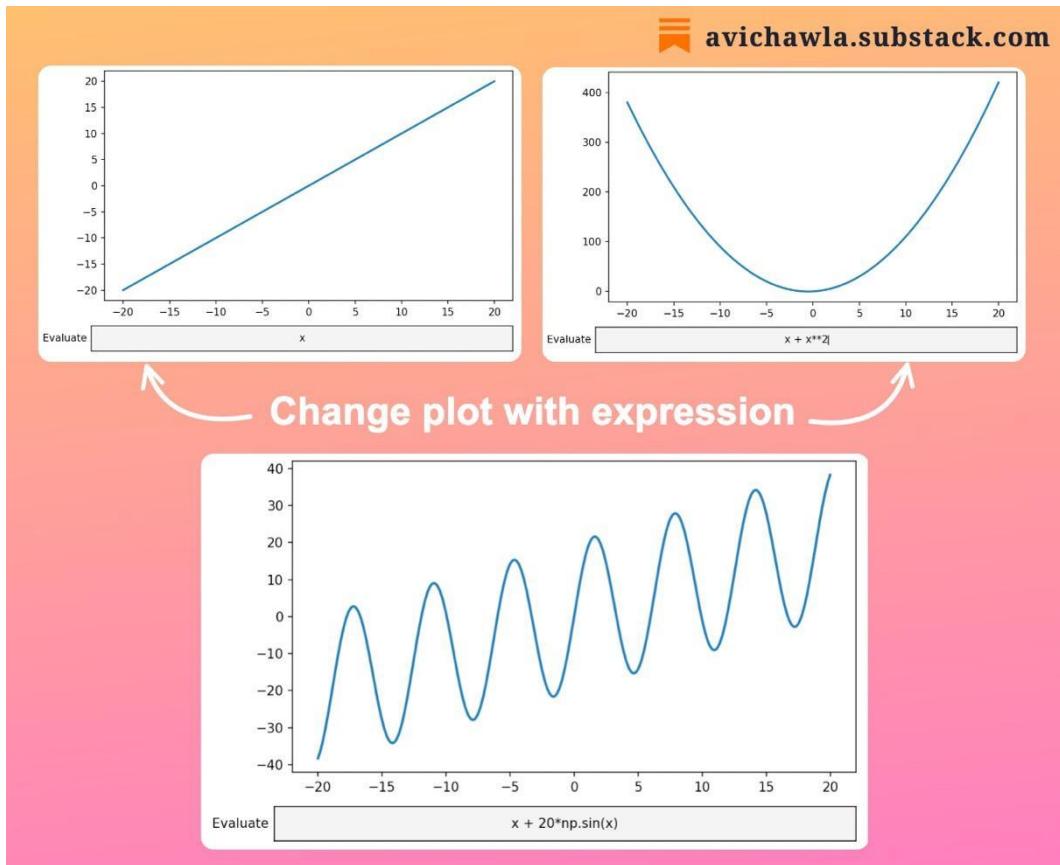
Instead, Python invokes the `__new__()` method first to create a new object and allocate memory to it. But how is that useful, you may wonder? There are many reasons.

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.

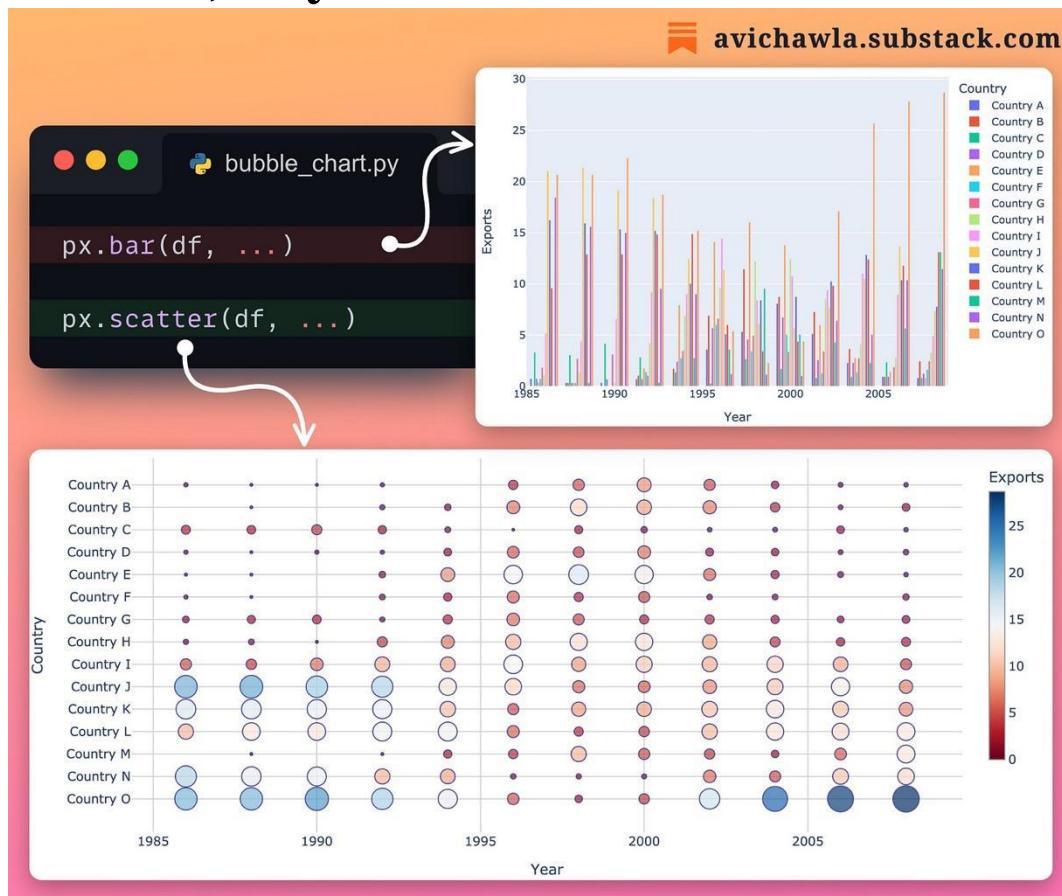
By default, Matplotlib uses the **inline** mode, which renders static plots. However, with the **%matplotlib widget** magic command, you can enable interactive backend for Matplotlib plots.

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)!

The image shows a Substack post by avichawla.substack.com. It contains two screenshots of a terminal window.

Top Screenshot: A dark terminal window showing Python code. The code creates a dictionary and a list, then tries to add the list to the dictionary as a key. It fails with a `TypeError: unhashable type: 'list'`.

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

Bottom Screenshot: A light-colored terminal window showing a workaround. It defines a new class `MyList` that inherits from `list` and implements the `__hash__` method to return 0. Then it creates an instance of `MyList` containing the list [1, 2, 3], adds it to a dictionary, and prints the dictionary to show the key is now hashable.

```
## 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}
```

Annotations on the left side of the image point to the word "hashable" in the first screenshot and the word "list" in the second screenshot. Annotations on the right side point to the word "unhashable" in the first screenshot and the word "list" in the second screenshot.

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.

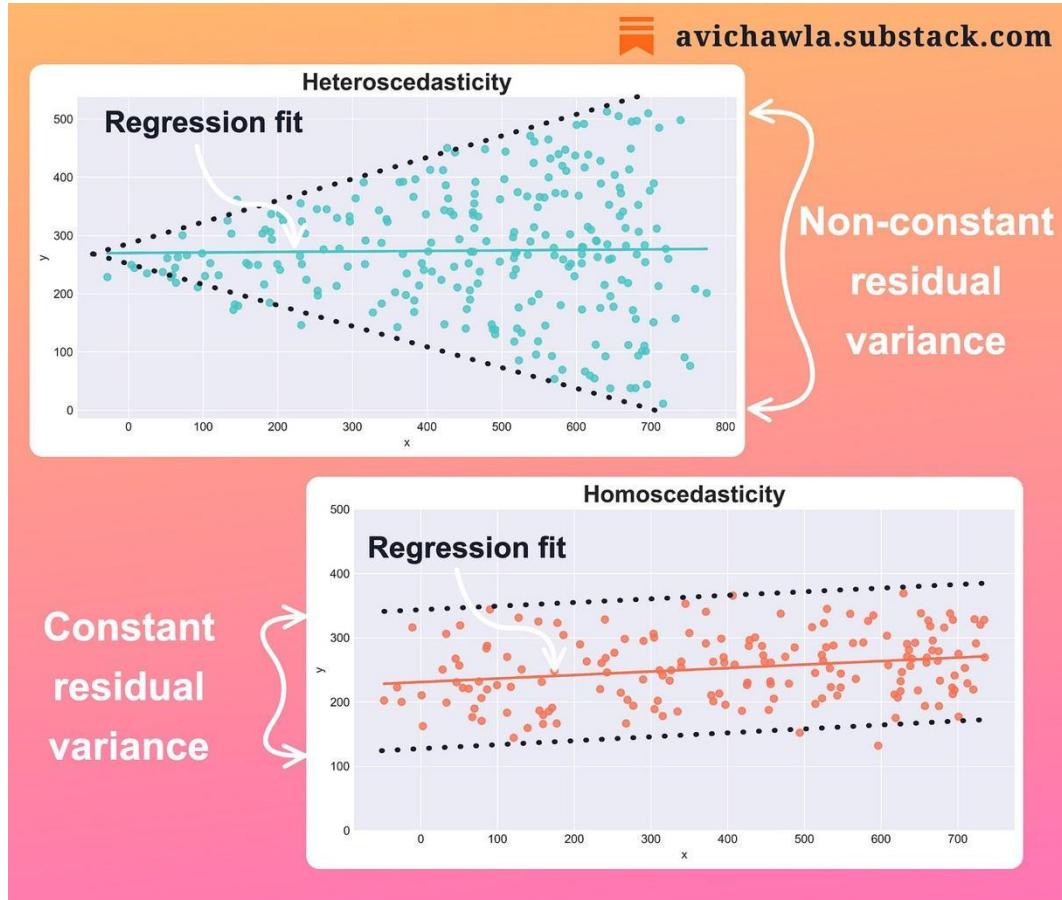
While classes of `int`, `str`, `tuple`, `frozenset`, etc. implement the `__hash__` method, it is missing from the `list` class. That is why we cannot add a list as a dictionary's key.

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.

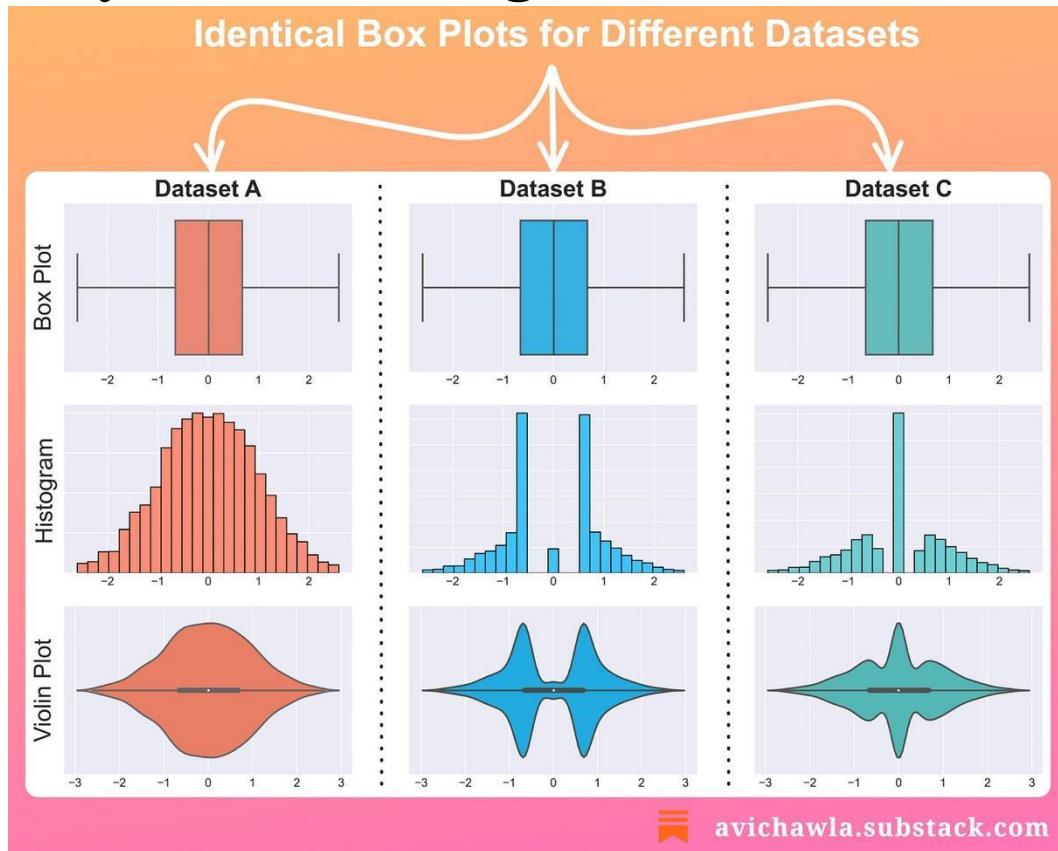
Here are some of them that will make you fall in love with Python and its versatility (even more).

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.

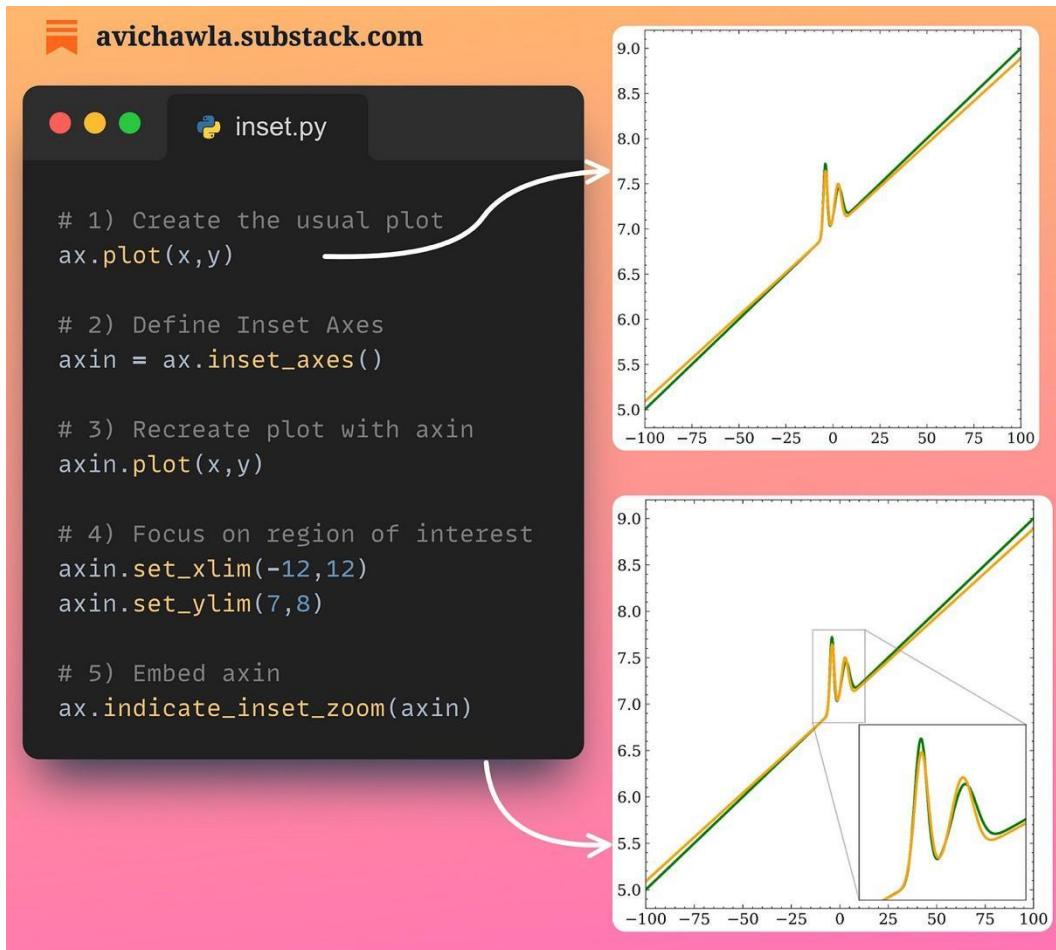
Thus, two different datasets with similar five values will produce identical box plots. This, at times, can be misleading and one may draw wrong conclusions.

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](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.indicate_inset_zoom.html).



The Pandas DataFrame Extension Every Data Scientist Has Been Waiting For

The screenshot shows a Jupyter Notebook interface with the title "Kanaries/pygwalker". The code cell contains:

```
[2]: import pandas as pd  
df = pd.read_csv('../bike_sharing_dc.csv', parse_dates=['date'])  
import pygwalker as pgw  
pgw.walk(df, hidebatisSourceConfig=True, vegaTheme='gv')
```

The visualization tab displays a stacked area chart with the X-axis labeled "hour" (0 to 24) and the Y-axis labeled "count". The legend indicates four categories: "season" (fall, spring, summer, winter) and "work yes or not" (0, 1). The chart shows two main peaks: one around 12 PM (labeled "hour") with a value of approximately 60,000, and another around 6 PM with a value of approximately 200,000. The data is color-coded by season: fall (blue), spring (green), summer (yellow), and winter (orange).

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

The screenshot shows the Xonsh shell interface with four panels:

- Run shell commands:**

```
$ cat file.txt ✓  
$ cd Desktop ✓
```
- Use Python:**

```
$ import pandas as pd ✓  
$ my_list = [1,2,3] ✓
```
- Shell commands with Python:**

```
$ for i in range(5):  
    echo @i ✓
```
- Shell commands with Python:**

```
$ var = 'he' + 'llo' ✓  
$ echo @var | grep "ll" ✓
```

A red arrow points from the text "Shell commands with Python" to the third and fourth panels.

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

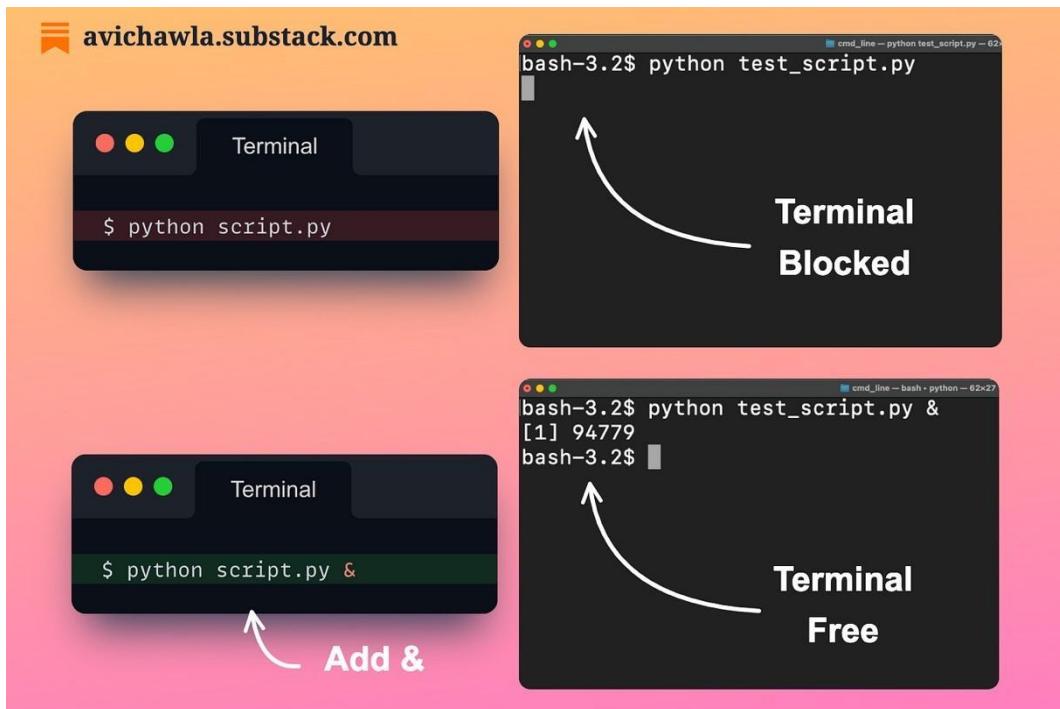
As a result, one has to open multiple terminals or switch back and forth between them in the same terminal.

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.

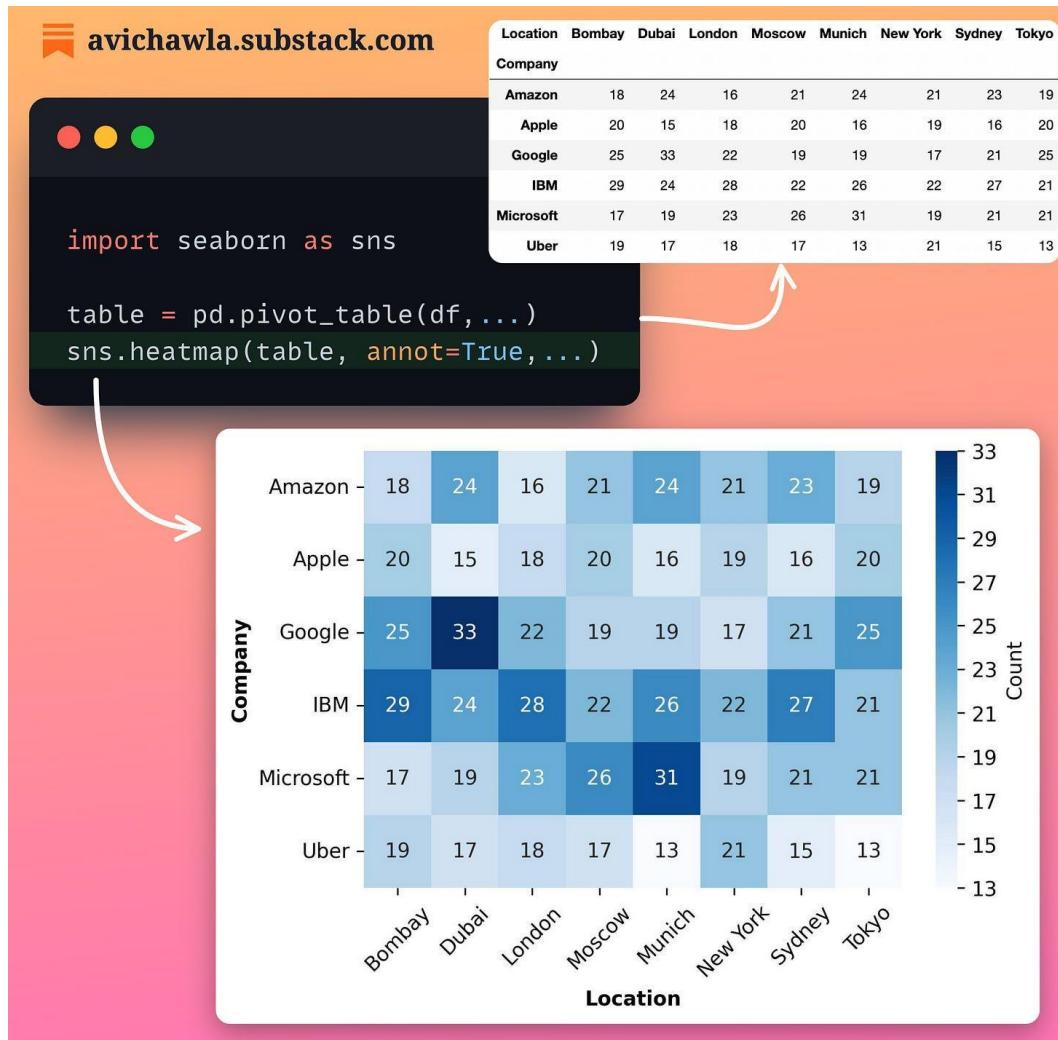
However, if you add '**&**' at the end of the command, the program will run in the background and instantly free the terminal.

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 Substack post by [avichawla.substack.com](#). It contains Python code defining a class `MyClass` with three types of attributes: public, protected, and private. Below the code, a terminal window shows the creation of an object and the access to these attributes. To the right, arrows point from the attribute names to their descriptions of accessibility.

```
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"
```

<pre>my_obj = MyClass()</pre>	<pre>>>> my_obj.public_attr</pre>	Public member accessible
	<pre>"I'm public"</pre>	
	<pre>>>> my_obj._protected_attr</pre>	Protected member accessible
	<pre>"I'm protected"</pre>	
	<pre>>>> my_obj._MyClass__private_attr</pre>	Private member accessible with name mangling
	<pre>"I'm private"</pre>	

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 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

The diagram illustrates the limitation of Python's native function overloading and how the `@dispatch` decorator from the `multipledispatch` library enables it.

Top Panel: Shows a Python interpreter window with the URL `avichawla.substack.com`. It contains the following code:

```
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'
```

A callout from this panel points to the text: "python interpreter only considers the latest definition of `add()` function".

Bottom Panel: Shows another Python interpreter window with the following code using the `@dispatch` decorator:

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

A callout from this panel points to the text: "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` → Hint: df.itterrows 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.itterrows()`
- In [5]:** `iter_df = df.itterrows()` → Hint: 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 [6]:** `df["new_col"] = df.apply(apply_func)` → Hint: 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 [7]:** `merged_df = pd.concat((df, df))` → Hint: After concatenation you have duplicated indices - pay attention
Line 1: `merged_df = pd.concat((df, df))`

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 Jupyter Notebook interface with two code cells. The top cell imports `bottleneck` and `numpy`, and creates a random array `arr`. The bottom section compares the run-times of various NumPy methods (`sum`, `mean`, `std`, `median`, `max`) against their corresponding `bottleneck` implementations. Arrows point from each NumPy method to its faster `bottleneck` counterpart.

NumPy Method	Bottleneck Method	Run-time	Performance Boost
<code>np.sum(arr)</code>	<code>bn.nansum(arr)</code>	<code>## Run-time: 870 μs</code>	<code>## Run-time: 33.9 μs (25x Faster)</code>
<code>np.mean(arr)</code>	<code>bn.nanmean(arr)</code>	<code>## Run-time: 477 μs</code>	<code>## Run-time: 21 μs (22x Faster)</code>
<code>np.std(arr)</code>	<code>bn.nanstd(arr)</code>	<code>## Run-time: 687 μs</code>	<code>## Run-time: 175 μs (4x Faster)</code>
<code>np.median(arr)</code>	<code>bn.nanmedian(arr)</code>	<code>## Run-time: 1.58 ms</code>	<code>## Run-time: 0.43 ms (4x Faster)</code>
<code>np.max(arr)</code>	<code>bn.nanmax(arr)</code>	<code>## Run-time: 1.26 ms</code>	<code>## Run-time: 0.46 ms (3x Faster)</code>

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

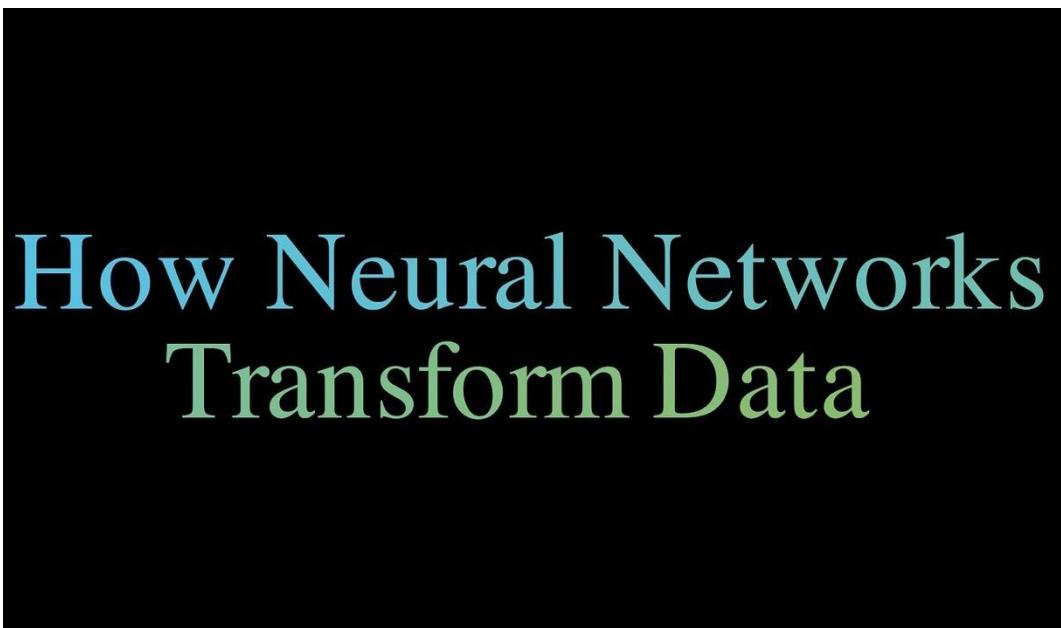
Bottleneck provides a suite of optimized implementations of NumPy methods.

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!

The diagram illustrates the refactoring process using Sourcery. It is divided into three main sections:

- Before Refactoring:** A screenshot of a Mac OS X desktop showing a terminal window titled "my_code.py" containing Python code. The code defines a function `is_special_number` that checks if a number is 7 or 18, returning True for either case and False otherwise.
- Refactoring:** A screenshot of a Mac OS X desktop showing a terminal window titled "Command Line" with the command `$ sourcery review --in-place my_code.py` entered.
- After Refactoring:** A screenshot of a Mac OS X desktop showing a terminal window titled "my_code.py" containing the refactored Python code. The function now returns the number directly if it is in the list [7, 18].

A white arrow points from the "Before Refactoring" section to the "Refactoring" section. Another white arrow points from the "Refactoring" section to the "After Refactoring" section. At the bottom left, there is a LinkedIn link: 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.

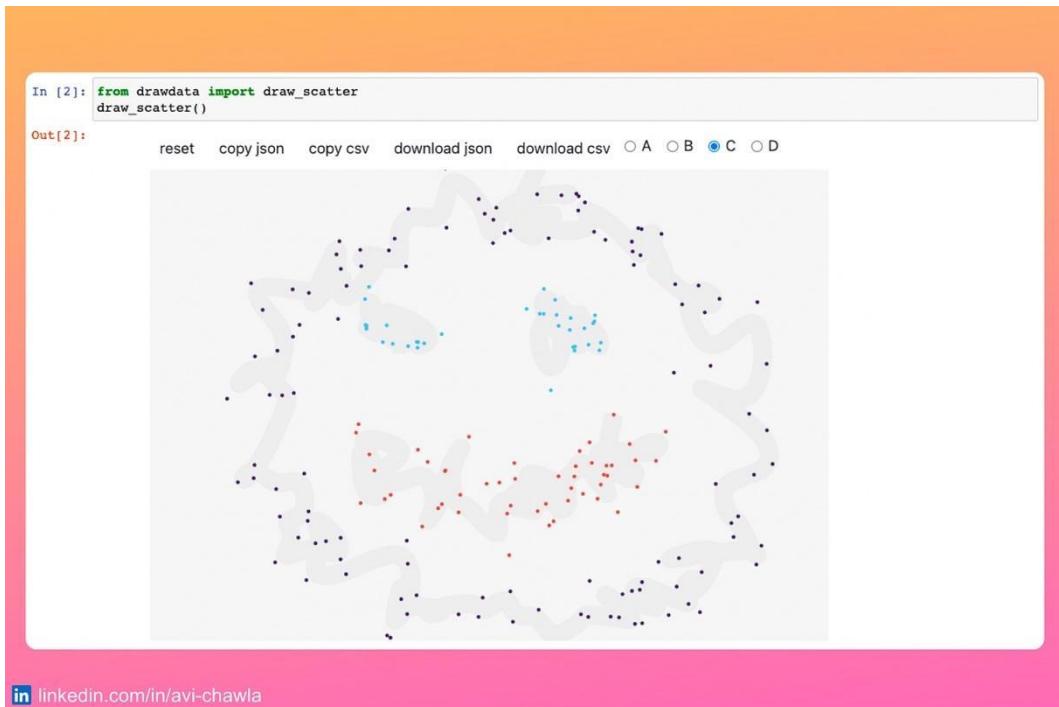
Instead, use Sourcery. It's an automated refactoring tool that makes your code elegant, concise, and Pythonic in no time.

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 slide compares two Python scripts for reading a 'file.parquet' Parquet file containing 32M rows.

Pandas Script:

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

fastparquet Script:

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

A white arrow points from the 'fastparquet.py' section towards the '5x Faster' text.

5x Faster

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

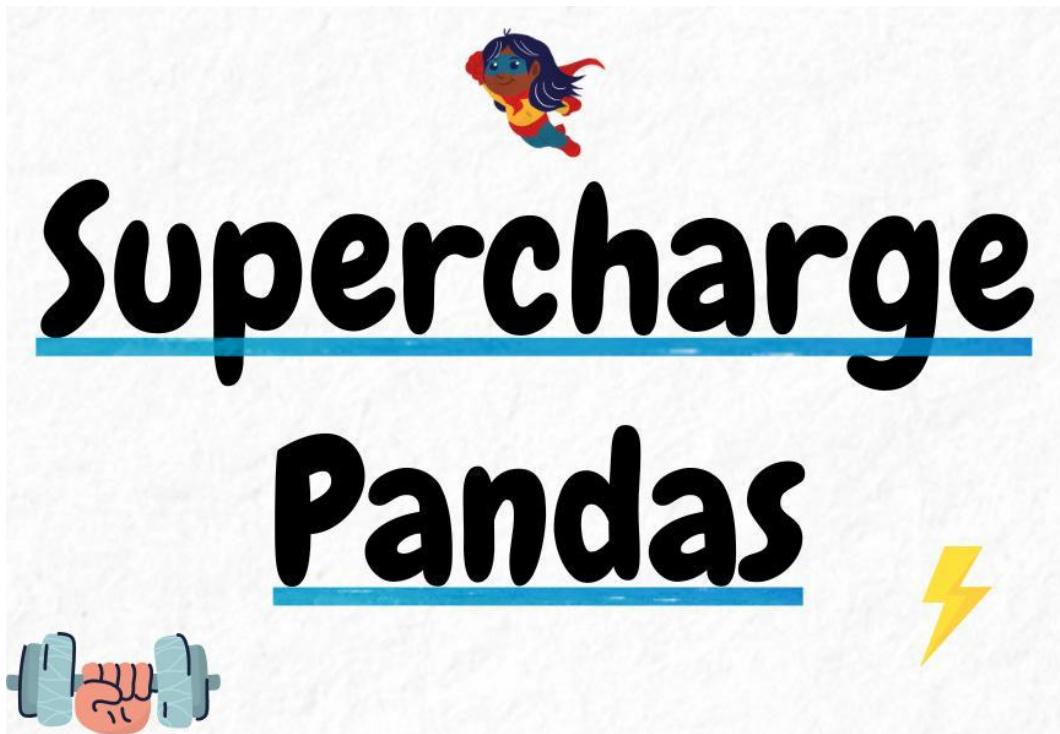
Dataframes are often stored in parquet files and read using Pandas' `read_parquet()` method.

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		Categories			
dataframe	Values	Column Type	Count						
Number of rows	1000	float64	3						
Number of columns	10	category	2						
		datetime64	2						
		int64	1						
		bool	1						
		string	1						
number									
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	
category									
column_name	NA	NA %	ordered	unique					
class	0	0	False	2					
location	1	0.1	False	5					
datetime									
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				
string									
column_name	NA	NA %	words per row		total words				
text	6	0.6			5.8	5800			
bool									
column_name	true	true rate			hist				
booly_col	520	0.52							

End

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

Add
decorator

Prints
warning

```
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.)
```

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
```

Method 1

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

```
petal_width <= 0.8
gini = 0.667
samples = 150
value = [50, 50, 50]
class = setosa

gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa

petal_width <= 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor

gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor

gini = 0.043
samples = 46
value = [0, 1, 45]
class = virginica
```

Method 2

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

```
--- 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.

```
1 from summarytools import dfSummary
2 dfSummary(iris_df)
```

No	Variable	Stats / Values	Freqs / (% of Valid)	Graph	Missing
1	sepal_length [float64]	Mean (sd) : 5.8 (0.8) min < med < max: 4.3 < 5.8 < 7.9 IQR (CV) : 1.3 (7.1)	35 distinct values		0 (0.0%)
2	sepal_width [float64]	Mean (sd) : 3.1 (0.4) min < med < max: 2.0 < 3.0 < 4.4 IQR (CV) : 0.5 (7.0)	23 distinct values		0 (0.0%)
3	petal_length [float64]	Mean (sd) : 3.8 (1.8) min < med < max: 1.0 < 4.3 < 6.9 IQR (CV) : 3.5 (2.1)	43 distinct values		0 (0.0%)
4	petal_width [float64]	Mean (sd) : 1.2 (0.8) min < med < max: 0.1 < 1.3 < 2.5 IQR (CV) : 1.5 (1.6)	22 distinct values		0 (0.0%)
5	species [object]	1. setosa 2. versicolor 3. virginica	50 (33.3%) 50 (33.3%) 50 (33.3%)		0 (0.0%)

in linkedin.com/in/avi-chawla

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

Terminal

```
$ nbgrep "import os" ./
```

Search in all notebooks

Benchmark.ipynb	:	cell 3:line 1 : import os
modin.ipynb	:	cell 1:line 3 : import os
kmeans.ipynb	:	cell 3:line 1 : import os

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

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](https://github.com/avachawla/nbcommands).



F-strings Are Much More Versatile Than You Think

The diagram illustrates two sections of code examples for f-strings:

Formatting:

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

Converting:

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

A LinkedIn link is also present: linkedin.com/in/avi-chawla.

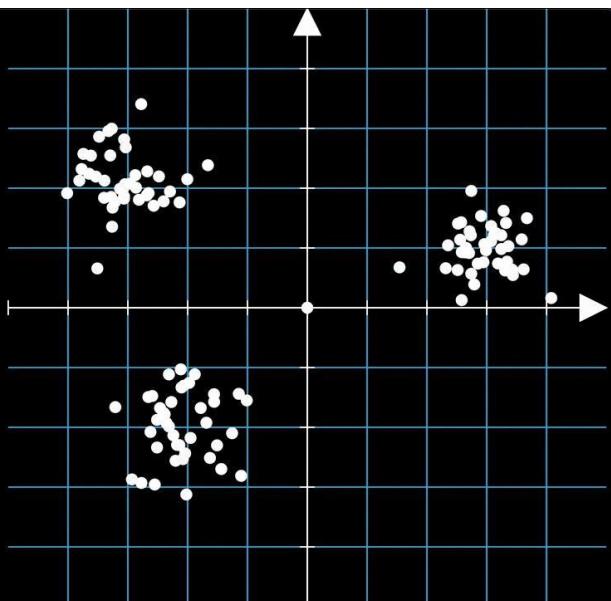
Here are 6 lesser-known ways to format/convert a number using f-strings. What is your favorite f-string hack?



Is This The Best Animated Guide To KMeans Ever?

K-Means Clustering Algorithm

By: Avi Chawla
avichawla.substack.com



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.

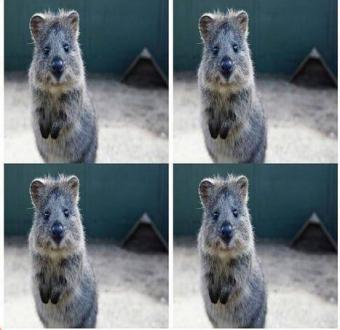
I created a beautiful animation using Manim to help you build an intuitive understanding of the algorithm.

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



An Effective Yet Underrated Technique To Improve Model Performance

Original Images

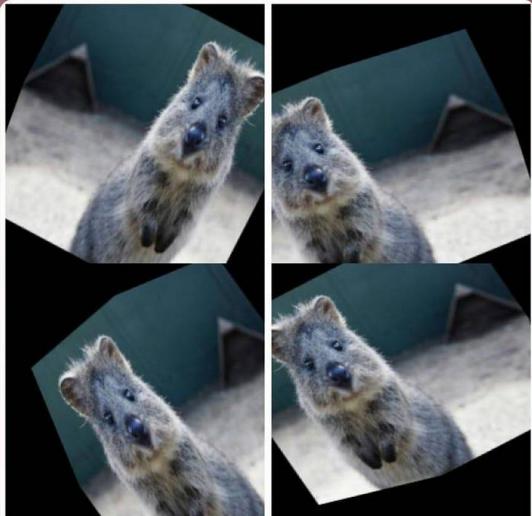


```
import imgaug.augmenters as iaa

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

images_aug = seq(images=images)
```

Augmented Images



[in](https://www.linkedin.com/in/avi-chawla) [linkedin.com/in/avi-chawla](https://www.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  ooooooo  oooooo
29|      ooooooo  ooooooo  oooooo
26|      ooooooo  ooooooo  oooooo
23|      ooooooo  ooooooo  oooooo
20|      ooooooo  ooooooo  oooooo
17|      ooooooo  ooooooo  oooooo  o
15|      ooooooo  ooooooo  oooooo  o
12|      ooooooo  ooooooo  oooooo  o
9|      ooooooo  ooooooo  oooooo  oooooo
6|      oo  ooooooo  ooooooo  oooooo  oooooo
3|      o  o  ooooooo  ooooooo  oooooo  oooooo
1|      o  o  ooooooo  ooooooo  oooooo  oooooo  o
```

 linkedin.com/in/avi-chawla

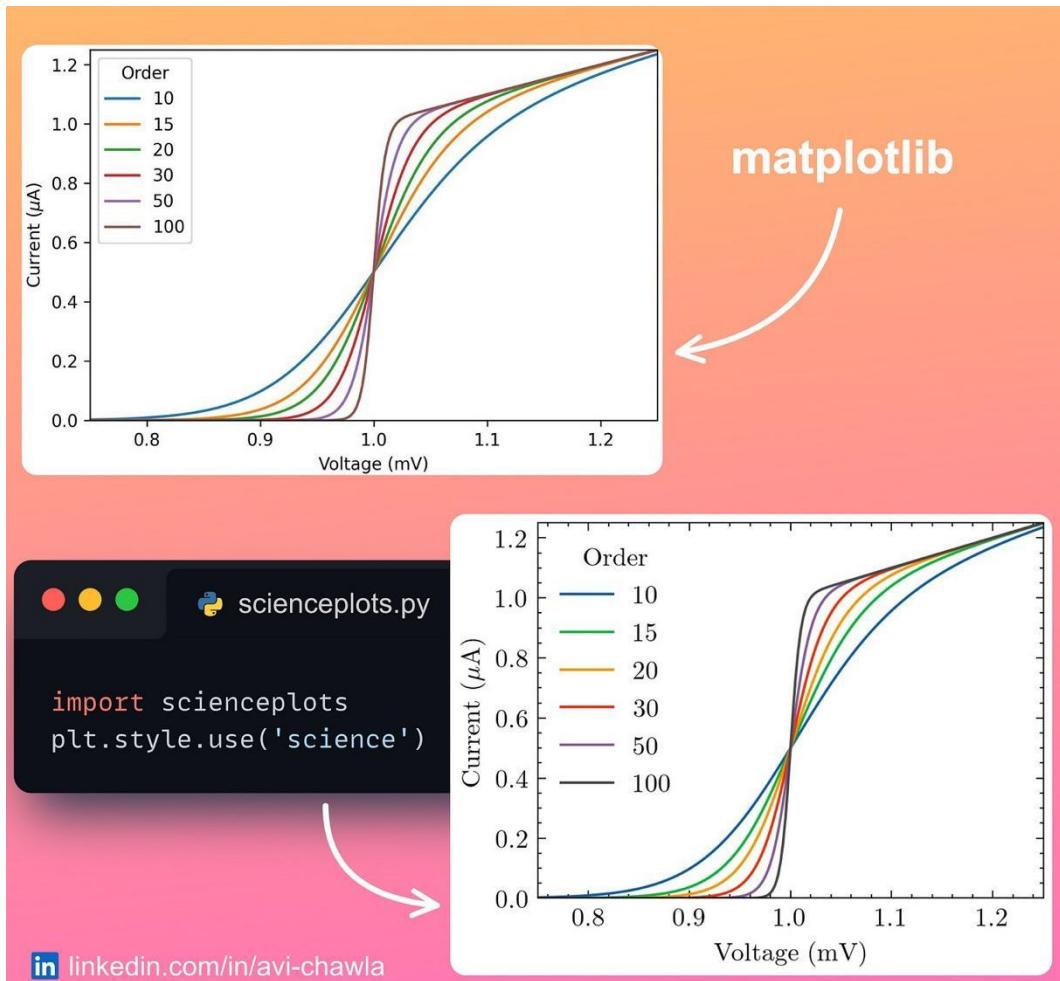
Visualizing data can get tough when you don't have access to a GUI. But here's what can help.

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](https://github.com/sciencedata/plots).



37 Hidden Python Libraries That Are Absolute Gems



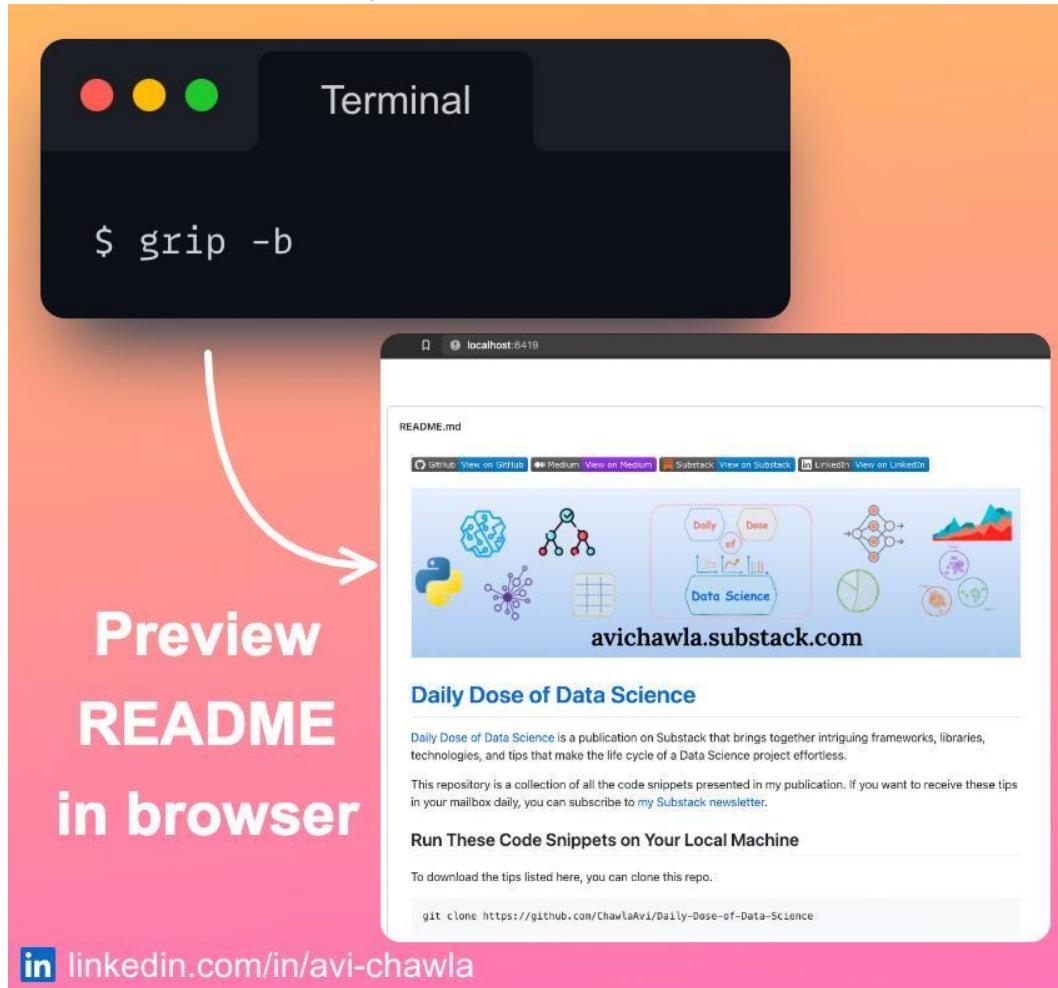
I reviewed 1,000+ Python libraries and discovered these hidden gems I never knew even existed.

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?

```
Std-dev.py
```

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

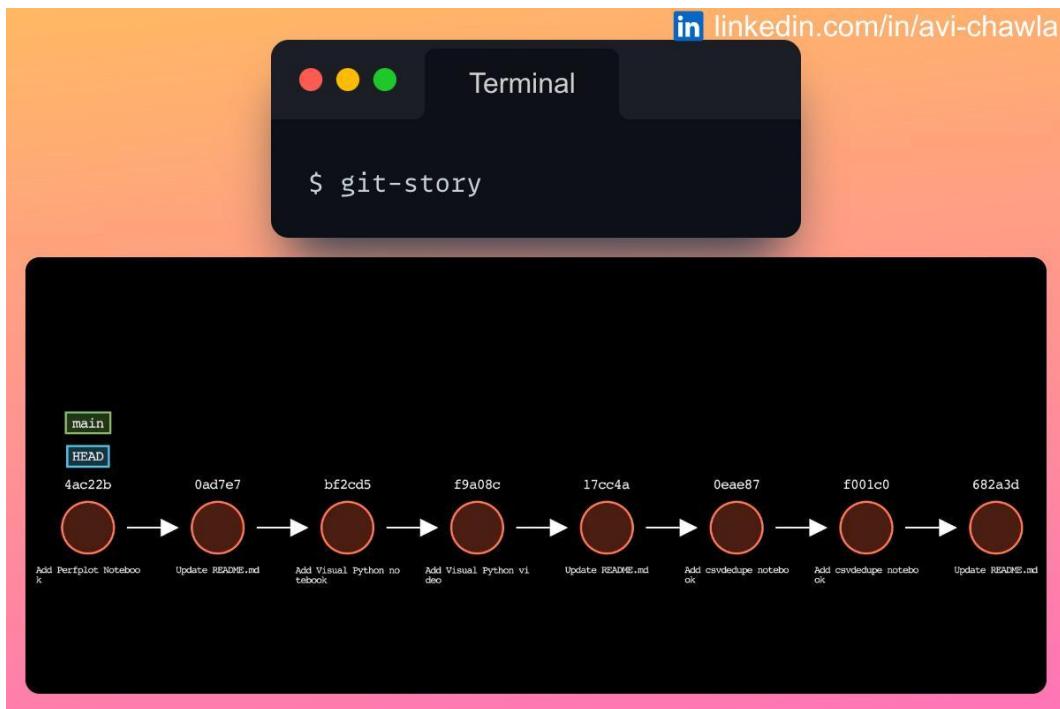
Thus, to generate an unbiased estimate, it uses $(n-1)$ as the dividing factor instead of n . In statistics, this is also known as Bessel's correction.

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.

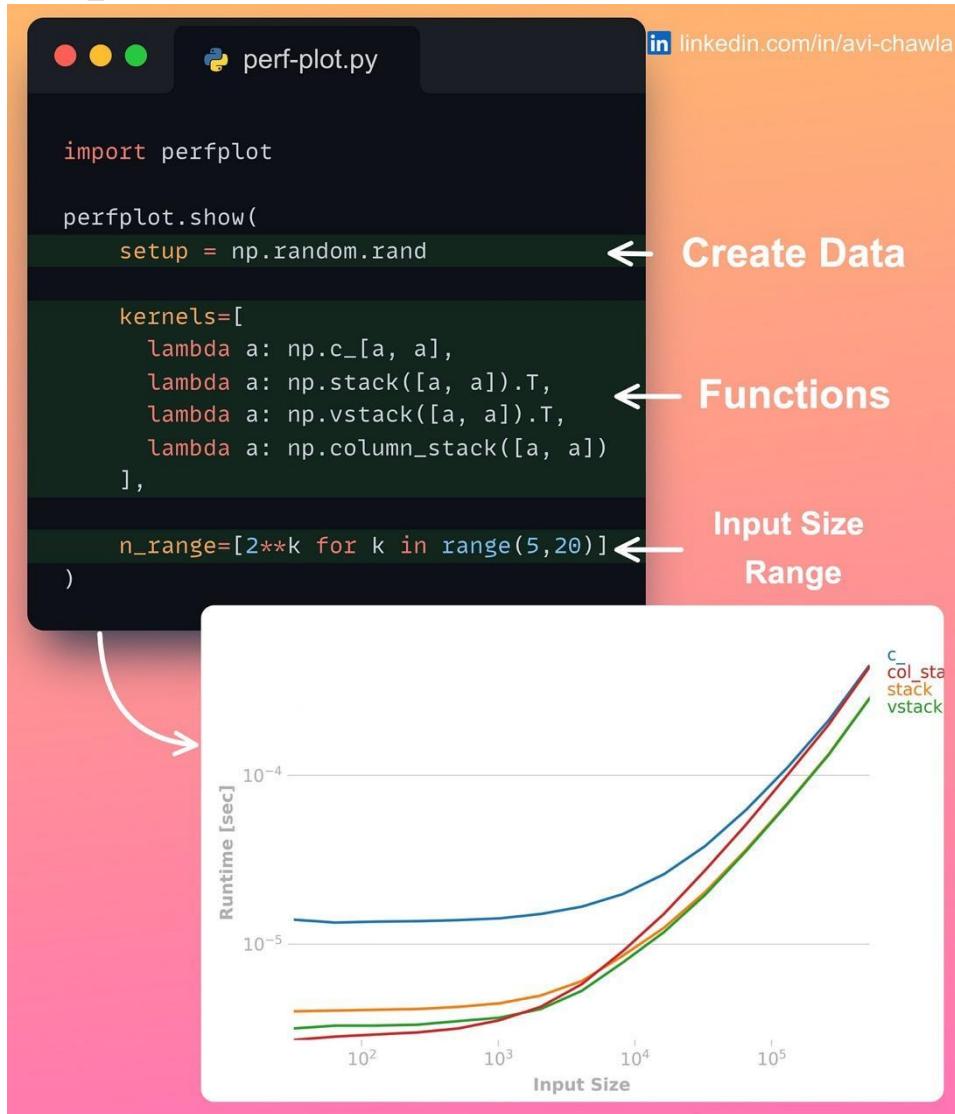
It generates a video that depicts the commits, branches, merges, HEAD commit, and many more. Find more info in the comments.

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

The screenshot shows a Jupyter notebook interface with three code cells:

```
In [1]: # Visual Python: Data Analysis > Import
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

In [2]: # Visual Python: Data Analysis > File
df = pd.read_csv('./dummy_data.csv')
df

...[redacted]

In [3]: # Visual Python: Visualization > Plotly
fig = px.scatter(df, x='Employee_Rating', y='Employee_Salary', color='Employment_Status')
fig.show()
```

A scatter plot titled "Employee_Salary" is displayed, showing a positive correlation between Employee_Rating and Employee_Salary. The legend indicates two categories: Intern (blue dots) and Full Time (red dots). The x-axis ranges from 0 to 5, and the y-axis ranges from 0 to 60k.

The right side of the interface features a "Visual Python" sidebar with various tools categorized into sections: Logic, Data Analysis, Visualization, and Machine Learning. The "Data Analysis" section includes buttons for Import, File, Variable, Shapes, Frame, Subplot, Instance, Groupby, Bind, Reshape, Markdown, and PDF. The "Visualization" section includes buttons for Chart Style, Pandas Plot, Matplotlib, Seaborn, Plotly, and WordCloud. The "Machine Learning" section includes buttons for Data Sets, Data Split, Data Prep, AutoML, Regressor, Classifier, Clustering, Dimension, Fit/Predict, Model Info, and Evaluation.

LinkedIn link: linkedin.com/in/avi-chawla

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

 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	0	Daniel	Lopez	719 Greene St. East Rhonda
1	0	Daniel	nan	719 Green Street East Rhoda
2	1	Alan	Martin	982 Carol Harbors Apart.
3	1	Alan Martin	nan	982 Carol Aparments
4	2	Philip	Owens	2578 Banks Ford
5	3	Shannon	White	USCGC Molina (150)082-7982
6	4	Julia	Anderson	09162 Mason Mnts.
7	4	Juliya	Anderrson	9162 Mason Street Mountain

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.

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")`

Print CSV Show 10 entries Search: andrade

	Name	Company_Name	Employee_Job_Title	Employee_City	Employee_Country
39	William Stein	Andrade LLC	Energy manager	North Melissafurt	Mali
585	Tyler Rasmussen	Andrade LLC	Equities trader	West Jamesview	Honduras
823	Tiffany Bailey	Andrade LLC	Energy manager	Aliciafort	Panama
411	Thomas Mullen	Andrade LLC	Make	West Jamesview	Niue
259	Steven Navarro	Andrade LLC	Patent examiner	West Jamesview	Gibraltar
32	Steven Burgess	Andrade LLC	Optometrist	New Cindychester	Saint Kitts and Nevis
166	Shelley Ramirez	Andrade LLC	Sales promotion account executive	New Russelton	Colombia
949	Sandra Gonzalez	Andrade LLC	Radiographer, therapeutic	North Melissafurt	Macedonia
426	Robert Reed	Andrade LLC	Sales promotion account executive	West Jamesview	British Virgin Islands
274	Robert Bryan	Andrade LLC	Trading standards officer	North Melissafurt	French Polynesia

Showing 1 to 10 of 69 entries (filtered from 1,000 total entries) Previous 1 2 3 4 5 6 7 Next

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

[in 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

The image shows a Mac desktop with two terminal windows side-by-side. The left window displays Python code for generating a list of pairs (a, b) where (a+b) % 11 == 0. The right window shows the same code using Numba's `@njit` decorator. A callout arrow from the text "≈33x Faster" points from the left window to the right window.

Left Terminal (Without Numba):

```
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
```

Right Terminal (With Numba):

```
from numba import njit

@njit
def func_with_numba():
    # same code

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

≈33x Faster

[in](https://www.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

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

```
from wordcloud import WordCloud  
  
wc = WordCloud().generate(text)
```

To the right of the code is a generated word cloud image. The image is a cloud shape filled with various words related to Python. The most prominent words are "Python", "language", "programming", and "use". The words are colored in shades of blue, green, and white against a black background.

The next cell in the notebook contains the following code:

```
from PIL import Image  
  
mask = Image.open("pylogo.png")  
  
wc = WordCloud(mask=mask,  
               contour_width=4,  
               ...)  
wc.generate(text)
```

To the right of this code is a more visually appealing word cloud. This one is shaped like the Python logo (a stylized 'P' and 'y'). The words are colored in shades of yellow, orange, and white against a white background. The word "Python" is very large and central, with "language" and "programming" being other major components.

in [linkedin.com/in/avi-chawla](https://www.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 word cloud will take its shape and appear fancier.

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



How to Encode Categorical Features With Many Categories?

```
import category_encoders as ce

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

enc.fit_transform(data["class"])

class_0    class_1    class_2
```

	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

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

} **Binary**

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

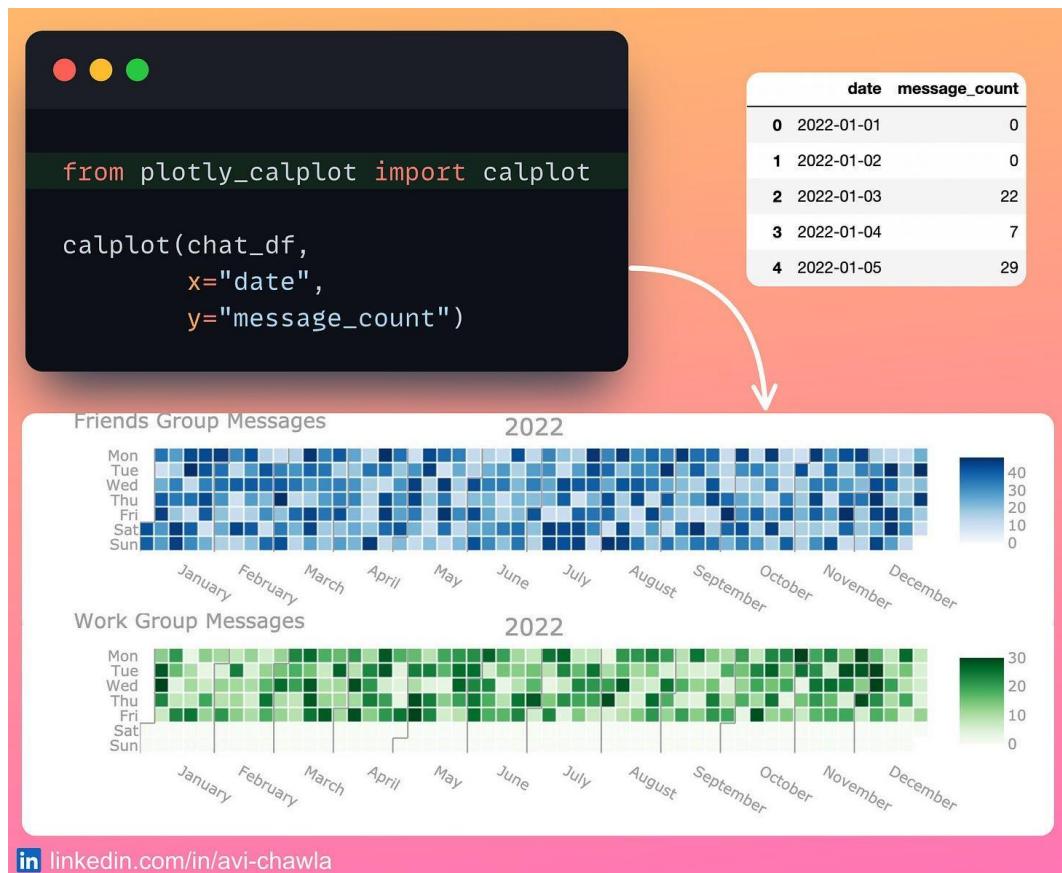
The category-encoders library provides a suite of encoders specifically for categorical variables. This makes it effortless to experiment with various encoding techniques.

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' and 'Correct clusters'.
Incorrect clusters: This section shows a scatter plot of data points forming two non-spherical clusters (Class A in blue, Class B in red). A green line represents the centroids assigned by KMeans, which do not correctly separate the two classes. The code for KMeans is shown as:

```
from sklearn import cluster
cluster.KMeans(2).fit(X)
```


Correct clusters: This section shows the same data points clustered correctly by DBSCAN, resulting in two distinct, well-separated circular clusters. The code for DBSCAN is shown as:

```
from sklearn import cluster
cluster.DBSCAN().fit(X)
```


A LinkedIn profile link is also present at the bottom left: 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
```

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

linkedin.com/in/avi-chawla

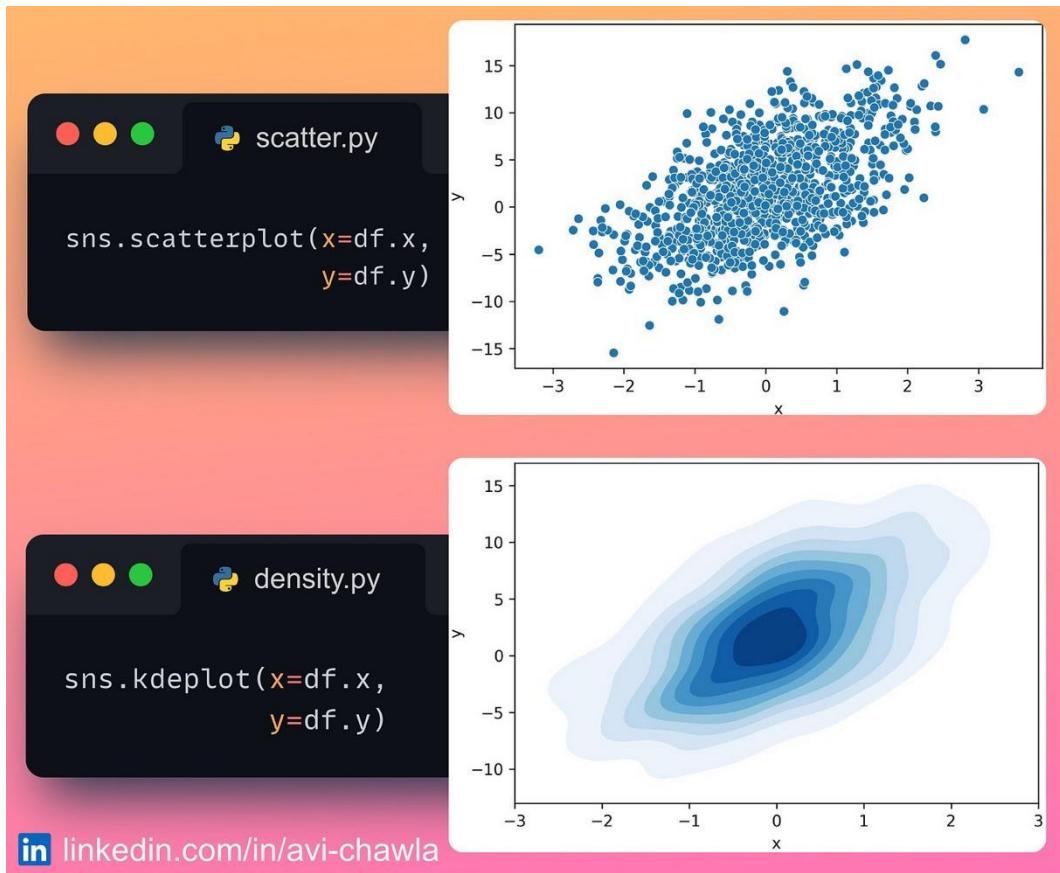
If you want to display python code and its output as LaTeX, try `latexify_py`. With this, you can print python code as a LaTeX expression and make your code more interpretable.

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

```
dummy_data.py

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

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](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html).



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
```





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

 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 Jupyter Notebook interface with three code cells and a data viewer window.

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)
```

Data Viewer:

The Pandas GUI window displays a table of data 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	Wendfort	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 Miquel
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

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

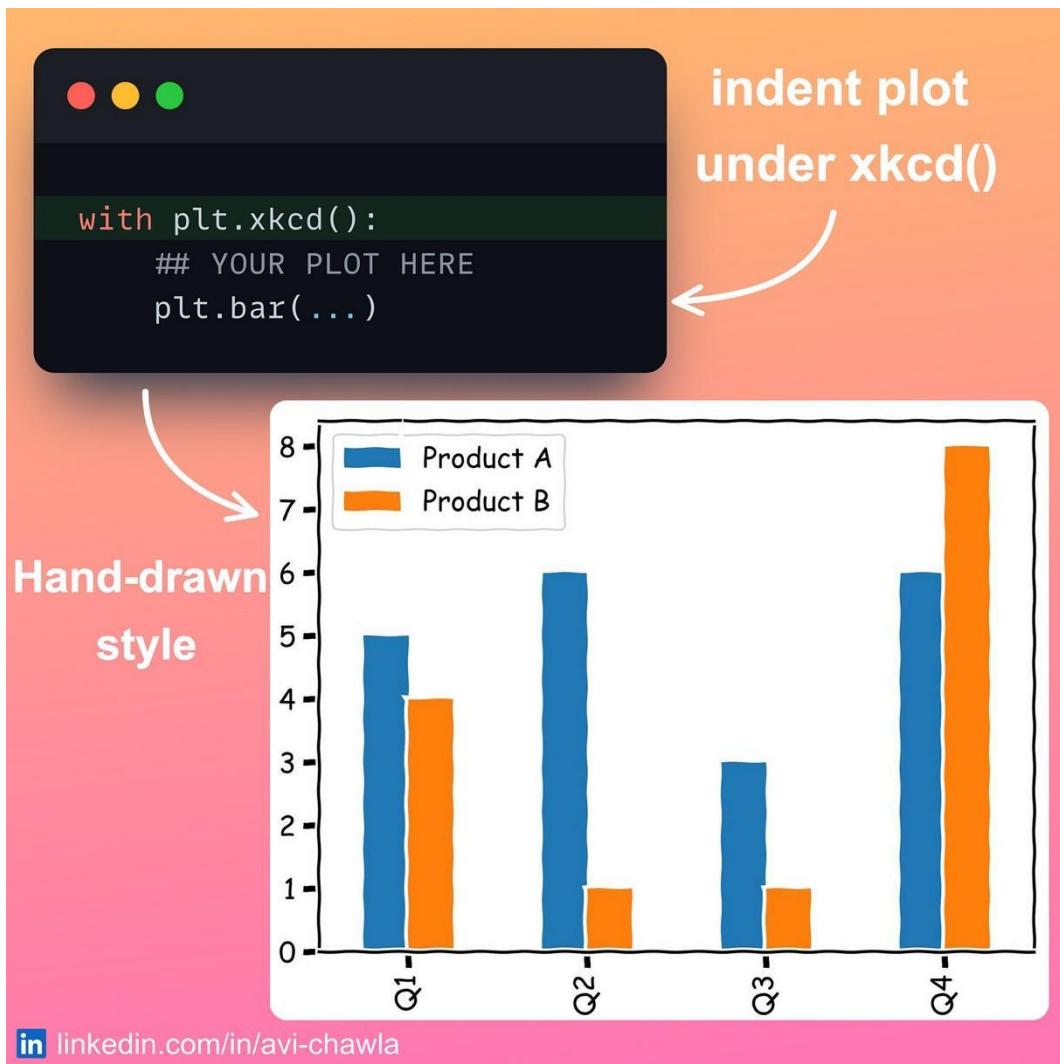
If you want to analyze your dataframe in a GUI-based application, try Pandas GUI. It provides an elegant GUI for viewing, filtering, sorting, describing tabular datasets, etc.

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 left window is titled "Pandas.py" and contains Python code for reading a 2M Row CSV file and concatenating it 20 times. The right window is titled "Modin.py" and contains similar code using modin.pandas instead of pandas. Arrows point from the "Import Statement" in the Modin code back to the Pandas code, and from the "7.1 sec" time in the Pandas code to the "0.1 sec" time in the Modin code. A large orange callout box labeled "7 GB Dataset" points to the dataset size in both code snippets.

```
Pandas.py:
```

```
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:
```

```
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)
```

in linkedin.com/in/avi-chawla

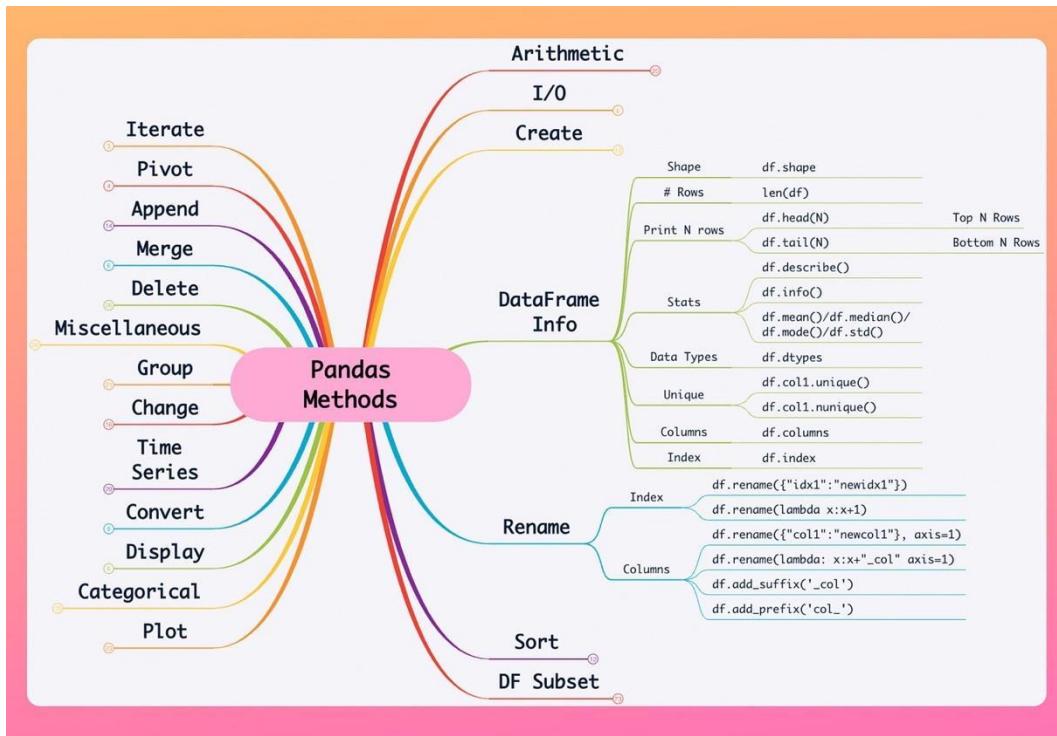
It is challenging to work on large datasets in Pandas. This, at times, requires plenty of optimization and can get tedious as the dataset grows further.

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

linkedin.com/in/avi-chawla

```
>>> s = Square(10)
>>> s.side # Getter
10

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

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 starts with a cell containing:

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

Then, another cell:

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

The output of this cell, `array([1, 2, 3])`, is highlighted with a green border and labeled **Out[2]:**.

Three arrows point from this output to subsequent cells:

- Arrow 1 points to cell **In [3]: _2**, which outputs `array([1, 2, 3])`.
- Arrow 2 points to cell **In [4]: Out[2]**, which also outputs `array([1, 2, 3])`.
- Arrow 3 points to cell **In [5]: _oh[2]**, which outputs `array([1, 2, 3])`.

A LinkedIn profile link is visible at the bottom left: linkedin.com/in/avi-chawla

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

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

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

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

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

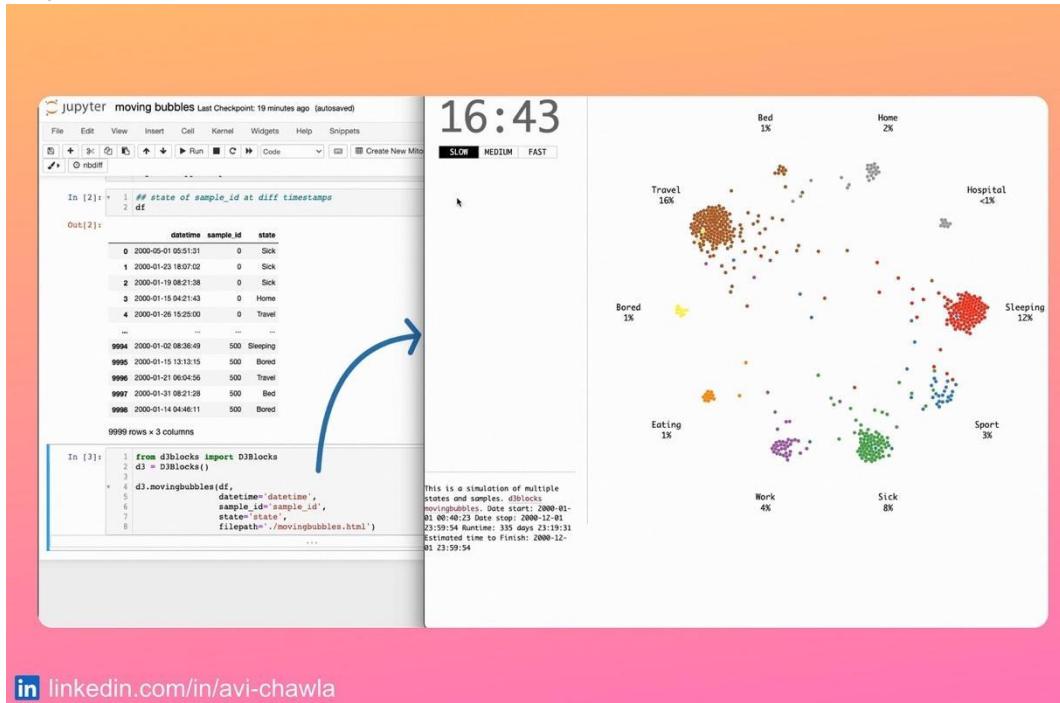
```
In [5]: _oh[2]
```

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



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

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 frame 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 Frame 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 assignment line `df["A"] = df.A.astype(np.int8)`.

LinkedIn Profile: 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://www.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.

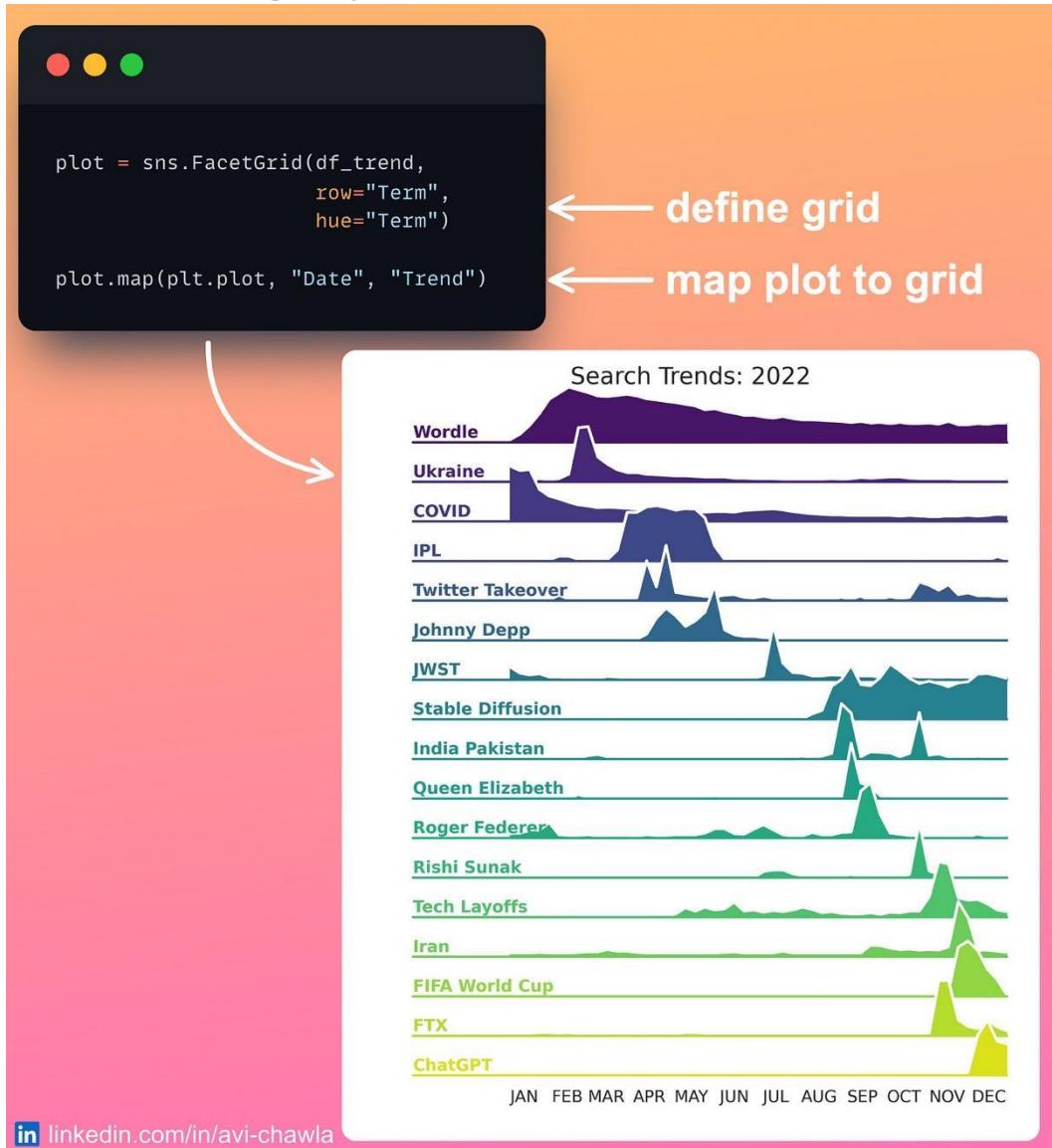
To invoke a method right before the interpreter is shutting down, decorate it with the `@atexit.register` decorator.

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.

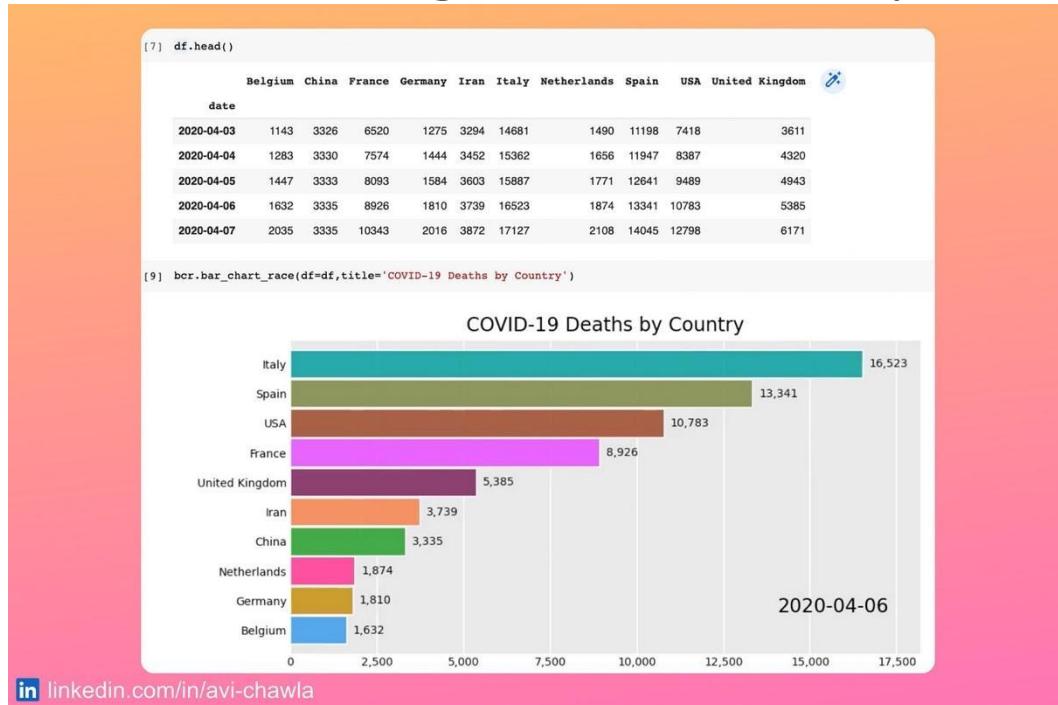
Instead, consider plotting the distribution across individual groups using FacetGrid. This allows you to compare the distributions of multiple groups side by side and see how they vary.

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

The screenshot shows two code cells side-by-side. The top cell, titled "Pandas Apply", contains the following Python code:

```
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
```

The bottom cell, titled "NumPy Select", contains the following Python code:

```
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
```

A large green arrow points from the Pandas code up to the NumPy code, with the text "Default" written below it. The Pandas code is labeled with "1.02 s" and "1.02 s ± 20.5 ms per loop". The NumPy code is labeled with "0.20 s" and "0.20 s ± 7.14 ms per loop".

On the right side of the slide, there is a small table with 5 rows and 5 columns, labeled "10^7 rows".

	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

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

Visit: playground.tensorflow.org

Tinker With a Neural Network Right Here in Your Browser.
Don't Worry, You Can't Break It. We Promise.

The screenshot shows the TensorFlow Playground interface. At the top, there are controls for 'Epoch' (set to 000,000), 'Learning rate' (0.1), 'Activation' (Tanh), 'Regularization' (None), 'Regularization rate' (0), and 'Problem type' (Classification). Below these are sections for 'DATA' (dataset selection, training/test ratio 90%, noise 15, batch size 10), 'FEATURES' (input selection: X_1 , X_2 , X_1^2 , X_2^2 , X_1X_2 , $\sin(X_1)$, $\sin(X_2)$), and 'OUTPUT' (test loss 0.358, training loss 0.320, scatter plot of data points colored by output values from -1 to 1). A note on the plot says 'This is the output from one neuron Hover to see it larger'. At the bottom left is a 'REGENERATE' button, and at the bottom right are checkboxes for 'Show test data' and 'Discretize output'.

linkedin.com/in/avi-chawla

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](http://playground.tensorflow.org).



What Are Class Methods and When To Use Them?

The screenshot shows a Python code editor window. The code defines a class `Rectangle` with an `__init__` method and a `@classmethod` named `from_square` which returns a `Rectangle` object with equal width and height. A white arrow points from the text "Define classmethod" to the `@classmethod` decorator in the code.

`class Rectangle:`
 `def __init__(self, width, height):`
 `self.width = width`
 `self.height = height`

 `@classmethod`
 `def from_square(cls, size):`
 `return Rectangle(size, size)`

create object using classmethod →

The screenshot shows a Python code editor window. It demonstrates creating a `Rectangle` object using the `from_square` classmethod with a parameter value of 5. The resulting object's width and height are printed, both outputting 5. A blue LinkedIn link is at the bottom left.

`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'. The code is as follows:

```
from sklearn.cluster import KMeans

kmeans = KMeans(8).fit(x_train)
# Training Time: 162s
```

To the right of the code, the text 'x_train shape: (500000, 1024)' is displayed. The bottom window is titled 'faiss.py' and also contains Python code for training a KMeans model using the Faiss library. The code is as follows:

```
import faiss

kmeans = faiss.Kmeans(d=1024, k=8)
kmeans.train(x_train)
# Training Time: 7.8s
```

A large green arrow points from the 'faiss.py' window towards the 'sklearn.py' window, with the text '≈20x Faster' written above the arrow.

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**

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 image shows a Jupyter Notebook interface with three code cells and a data preview area.

Code Cell 1:

```
def min_max(row):
    return max(row), min(row)
```

Code Cell 2:

```
>>> df.apply(min_max, axis = 1)
0  (3, 1)
1  (6, 3)
```

Data Preview:

	A	B	C
0	1	3	2
1	4	6	3

Annotations:

- An arrow points from the text "Pandas DataFrame" to the first code cell.
- An arrow points from the text "Pandas Series of Tuple" to the second code cell.

LinkedIn Profile:

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 in dark-themed code editors:

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
"""
```

A white arrow points from the text "list of dataclass objects" to the list of Point objects in the second snippet. Another white arrow points from the text "define dataclass" to the first snippet.

[in/avi-chawla](https://www.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 image shows two terminal windows side-by-side. The top window displays the following Python code:

```
from dataclasses import dataclass

@dataclass
class Student:
    name:str
    key:str

Jane = Student("Jane", "27HD")

print(Jane)
Student(name='Jane', key='27HD')
```

The output of this code is: `Student(name='Jane', key='27HD')`. An annotation with a curved arrow points from the text "Prints all attributes" to this output.

The bottom window displays the same code, but with one modification:

```
from dataclasses import dataclass, field

@dataclass
class Student:
    name:str
    key:str = field(repr = False)

Jane = Student("Jane", "27HD")

print(Jane)
Student(name='Jane')
```

The output of this code is: `Student(name='Jane')`. An annotation with a curved arrow points from the text "Attribute hidden in print" to the line `key:str = field(repr = False)`.

At the bottom left of the image is a LinkedIn link: 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
frozense

```
## frozenset
my_set = frozenset({1, 2, 3})

my_dict = {my_set: "A frozen set"}

my_dict[my_set]
"A frozen set"
```

in 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 environments side-by-side.

Top Environment (dot.py):

- Code:

```
>>> a:np.array # Shape: (a,b,c,d)
```

```
>>> b:np.array # Shape: (p,q,d,r)
```

```
>>> np.dot(a, b) # Shape: (a,b,c,p,q,r)
```
- Annotations:
 - Left arrow from the first code line to text: "a*b*c VECTORS of shape (d)"
 - Left arrow from the second code line to text: "p*q*r VECTORS of shape (d)"
 - Curved arrow from the bottom text to the third code line: "dot product of a*b*c and p*q*r VECTORS"

Bottom Environment (matmul.py):

- Annotations:
 - Right arrow from the text "a*b MATRICES of shape (c,d)" to the first code line in the environment.
 - Right arrow from the text "a*b MATRICES of shape (d,e)" to the second code line in the environment.
- Code:

```
>>> a:np.array # Shape: (a,b,c,d)
```

```
>>> b:np.array # Shape: (a,b,d,e)
```

```
>>> np.matmul(a, b) # Shape: (a,b,c,e)
```
- Annotation:
 - Curved arrow from the bottom text to the third code line: "Matrix product of a*b MATRICES"

Bottom-left corner: 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

The image shows two Jupyter notebook cells side-by-side against a background gradient from orange to pink.

Pandas Cell: Labeled "Filter-Pandas.ipynb". It contains the Python code: `df[df.city == "New Delhi"]`.

DuckDB Cell: Labeled "Filter-SQL.ipynb". It contains the SQL code: `%%sql select * from df where city = 'New Delhi';`.

LinkedIn Profile: A blue LinkedIn icon followed by the URL 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

The diagram illustrates the process of automated code refactoring using Sourcery. It consists of three main sections:

- Before Refactoring:** A screenshot of a Python code editor showing a function `set_my_var` with a complex if-else block:

```
def set_my_var(condition):
    if condition:
        my_var = 1
    else:
        my_var = 2

    return my_var
```
- Command Line:** A terminal window showing the command to run Sourcery's in-place review:

```
$ sourcery review --in-place my_code.py
```
- After Refactoring:** A screenshot of the code editor again, but now showing a simplified one-liner:

```
def set_my_var(condition):
    return 1 if condition else 2
```

Curved arrows indicate the flow from the original code to the refactored code, and from the command line to the refactored code.

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
```

[in 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

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 displays two screenshots of Python code in a terminal window. The top screenshot shows a file named `dotmap.py` with the following code:

```
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
```

To the right of this code, the text "Add using dot notation" is displayed with a curved arrow pointing from the text to the code.

The bottom screenshot shows a file named `dotmap-print.py` with the following code:

```
people pprint()

{'dave': {'english': 34, 'id': '12C'},
 'john': {'english': 45, 'id': '12A'},
 'mary': {'english': 49, 'id': '12B'}}
```

To the left of this code, the text "Pretty Print" is displayed with a curved arrow pointing from the text to the code.

At the bottom left of the image, there is a LinkedIn icon followed by the URL 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 image shows two Python code editors side-by-side. The left editor, titled 'my_functions.py', contains the following code:

```
__all__ = ["func1", "func2"]

def func1():
    return "Function 1"

def func2():
    return "Function 2"

def func3():
    return "Function 3"
```

A callout arrow from the text 'Specify __all__' points to the `__all__` assignment in the code. The right editor, titled 'module.py', contains the following code:

```
from my_functions import *

>>> func1()
"Function 1"

>>> func2()
"Function 2"

>>> func3()
NameError: name 'func3' is not defined
```

A callout arrow from the text 'Only func1 and func2 imported' points to the import statement in the code.

in 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 displays three separate code snippets, each enclosed in a dark rectangular box with a green rounded rectangle highlighting specific parameters. The background has a horizontal gradient from orange to pink.

- Top Snippet:** `pd.read_csv("data.csv", nrows = 10)`. To the right, the text "Read only first 10 rows" is displayed.
- Middle Snippet:** `pd.read_csv("data.csv", usecols = ["A", "C"])`. To the left, the text "Read specific columns" is displayed.
- Bottom Snippet:** `pd.read_csv("data.csv", skiprows = 10)` and `pd.read_csv("data.csv", skiprows = [1, 5])`. To the right, the text "Skip first 10 rows" is associated with the first snippet, and "Skip 1st row and 5th row" is associated with the second snippet. Arrows point from the text labels to their respective `skiprows` parameters.

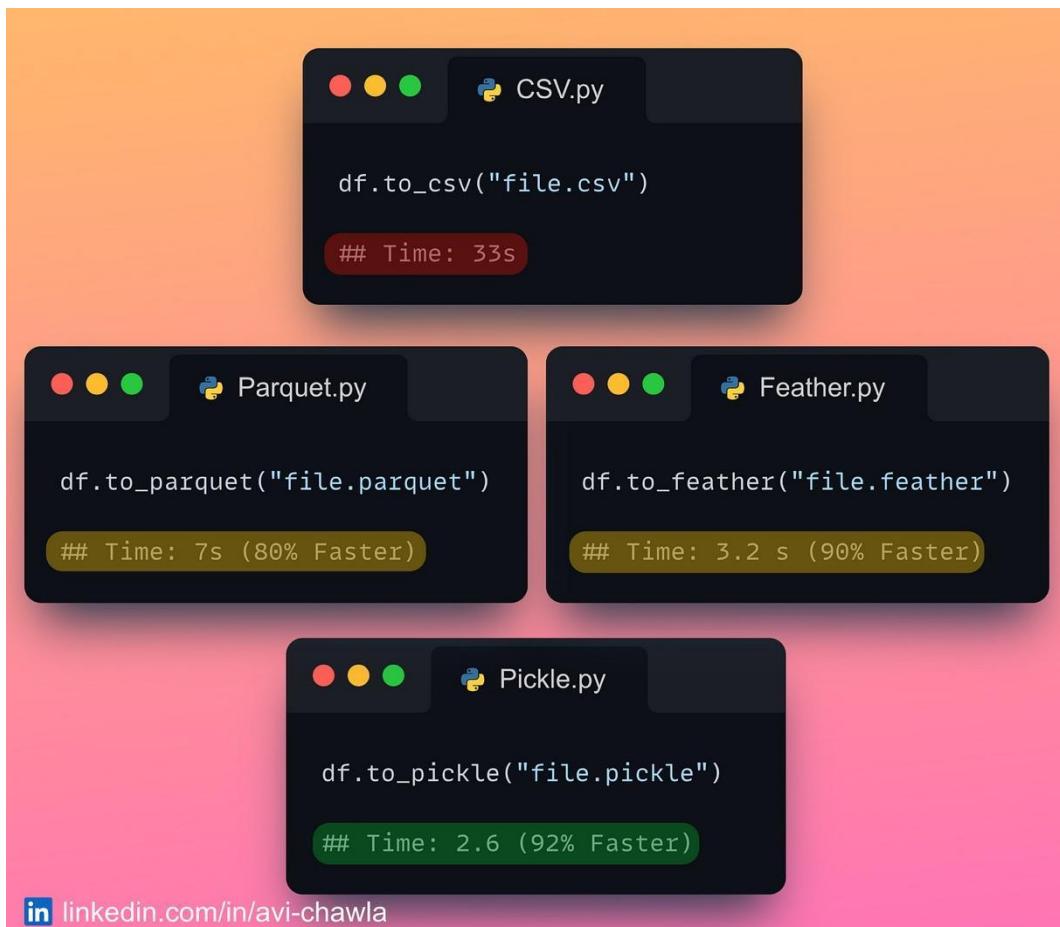
[in linkedin.com/in/avi-chawla](https://linkedin.com/in/avi-chawla)

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 a comparison between two terminal windows. The top window displays the Python code `py-snooper.py` with annotations. The line `@pysnooper.snoop()` is highlighted with a green background and labeled "Add Decorator". The bottom window shows the command `$ python py-snooper.py` followed by the debugging output. The output highlights variable values (`a = 9`, `b = 5`, `add = 14`, `sub = 4`) and the final return value (`(14, 4)`). Arrows from the text labels point to their corresponding elements in the code and output.

```
py-snooper.py
import pysnooper
@pysnooper.snoop()
def add_sub(a, b):
    add = a+b
    sub = a-b
    return (add, sub)
add_sub(9, 5)
```

```
$ python py-snooper.py
Starting var:.. a = 9
Starting var:.. b = 5
call line
New var:..... add = 14
line
New var:..... sub = 4
line
Return value:.. (14, 4)
```

in 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

```
pd.merge(name_df, rewards_df,
         on = "Cust_ID",
         how = "outer",
         indicator = True)
```

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

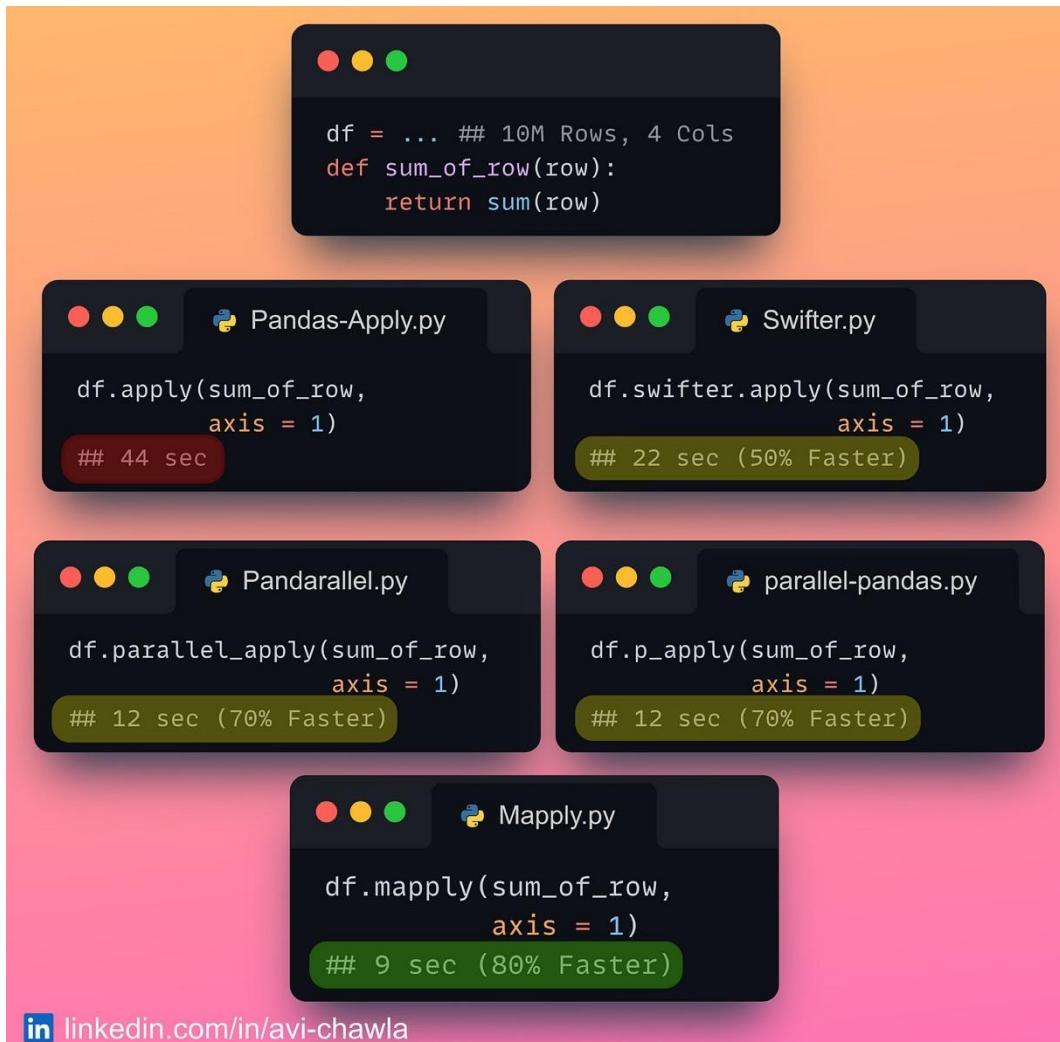
in linkedin.com/in/avi-chawla

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

batch = $X_{[i,:,:]}$

$\begin{array}{c} 764 \\ \text{---} \\ 10 \end{array}$ $\begin{array}{c} 10 \\ \text{---} \\ 20 \end{array}$ $\begin{array}{c} 164 \\ \text{---} \\ <\text{float32}> \end{array}$

$Y_{100} = \text{torch.matmul}(W_{764}, batch.T_{10}) + b_{100}$

$\begin{array}{c} 100 \\ \text{---} \\ 10 \end{array}$ $\begin{array}{c} 764 \\ \text{---} \\ <\text{float32}> \end{array}$ $\begin{array}{c} 10 \\ \text{---} \\ 764 \end{array}$ b_{100}_{1} $<\text{float32}>$

[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.

Only numbers

Summary of Array

linkedin.com/in/avi-chawla

```
>>> 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]])
```

```
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!
```

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 four terminal windows side-by-side, each with a dark background and light-colored text. The top-left window is titled 'fib.py' and contains Python code for calculating Fibonacci numbers. The top-right window is titled 'calc_pi.py' and contains Python code for approximating pi. The bottom-left window is titled 'python3.10.py' and shows benchmark results for Python 3.10. The bottom-right window is titled 'python3.11.py' and shows benchmark results for Python 3.11. Arrows point from the bottom-left window to the bottom-right window, indicating a comparison between the two versions.

Python Version	Function	Input	Time (ms)	Notes
Python 3.10	fib	30	260ms	
		40	32s	
	pi_approx	10**6	144ms	
Python 3.11	fib	30	97ms	(62% Faster)
		40	12s	(62% Faster)
	pi_approx	10**6	65ms	(55% Faster)

[linkedin.com/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

The screenshot shows a Jupyter Notebook interface. In the top cell (In [1]), the command `import pandas as pd` is run, followed by its execution time: "executed in 1.28s, finished 15:24:30 2022-12-06". In the second cell (In [2]), the user types `pd.DataFrame()`. By pressing Shift-Tab, a tooltip appears, displaying the function's signature and docstring. The signature is:

```
pd.DataFrame(  
    data=None,  
    index='Axes | None' = None,  
    columns='Axes | None' = None,  
    dtype='Dtype | None' = None,  
    copy='bool | None' = None,  
)
```

The docstring is:

```
Docstring:  
Two-dimensional size-mutable, potentially heterogeneous tabular data
```

At the bottom left of the notebook area, there is a LinkedIn link: 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 dark-themed application window titled "Pandas Profiling Report". At the top, there's a code editor containing Python code:

```
from pandas_profiling import ProfileReport

profile = ProfileReport(iris_data,
                       title="Pandas Profiling Report")

profile.to_widgets()
```

Below the code editor is the generated report interface. It has a tab bar with "Overview", "Variables", "Interactions", "Correlations", "Missing values", "Sample", and "Duplicate rows". The "Alerts (7)" tab is selected. The report lists several alerts:

- 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 report, there are two columns of status indicators:

Duplicates	High correlation
High correlation	High correlation
High correlation	High correlation
High correlation	High correlation
Uniform	

At the bottom left, it says "Report generated by YData". At the bottom center, there's a LinkedIn link: 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?

```
>>> a = 256
>>> b = 256

>>> a is b
True

>>> a = 257
>>> b = 257

>>> a is b
False

>>> a, b = 257, 257

>>> a is b
True
```

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





Make a Class Object Behave Like a Function

The diagram illustrates how to make a class object behave like a function by defining a `__call__` method. It consists of two main sections:

Top Section (Code Snippet):

```
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
```

Annotations:

- A callout bubble points to the `__call__` method definition with the text "define __call__ method".
- A callout bubble points to the **class object behaves like function** text with two arrows.

Bottom Section (Code Snippet):

```
f = Quadratic(1, 2, 3)

print(f(1)) # Output: 6
print(f(2)) # Output: 11

print(callable(f)) # Output: True
```

Annotation:

- A callout bubble points to the `print(f(1))` line with the text "class object behaves like function".

Link:

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



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 Python code editors side-by-side. The left editor, titled 'dump.py', contains the following 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)
```

An annotation 'Store 3 Variables' with a white arrow points to the 'with' block. The right editor, titled 'load.py', contains the following code:

```
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
```

An annotation 'Load Only First 2 Variables' with a white arrow points to the first two 'pickle.load' calls.

linkedin.com/in/avi-chawla

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.



DailyDoseofDS.com



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.

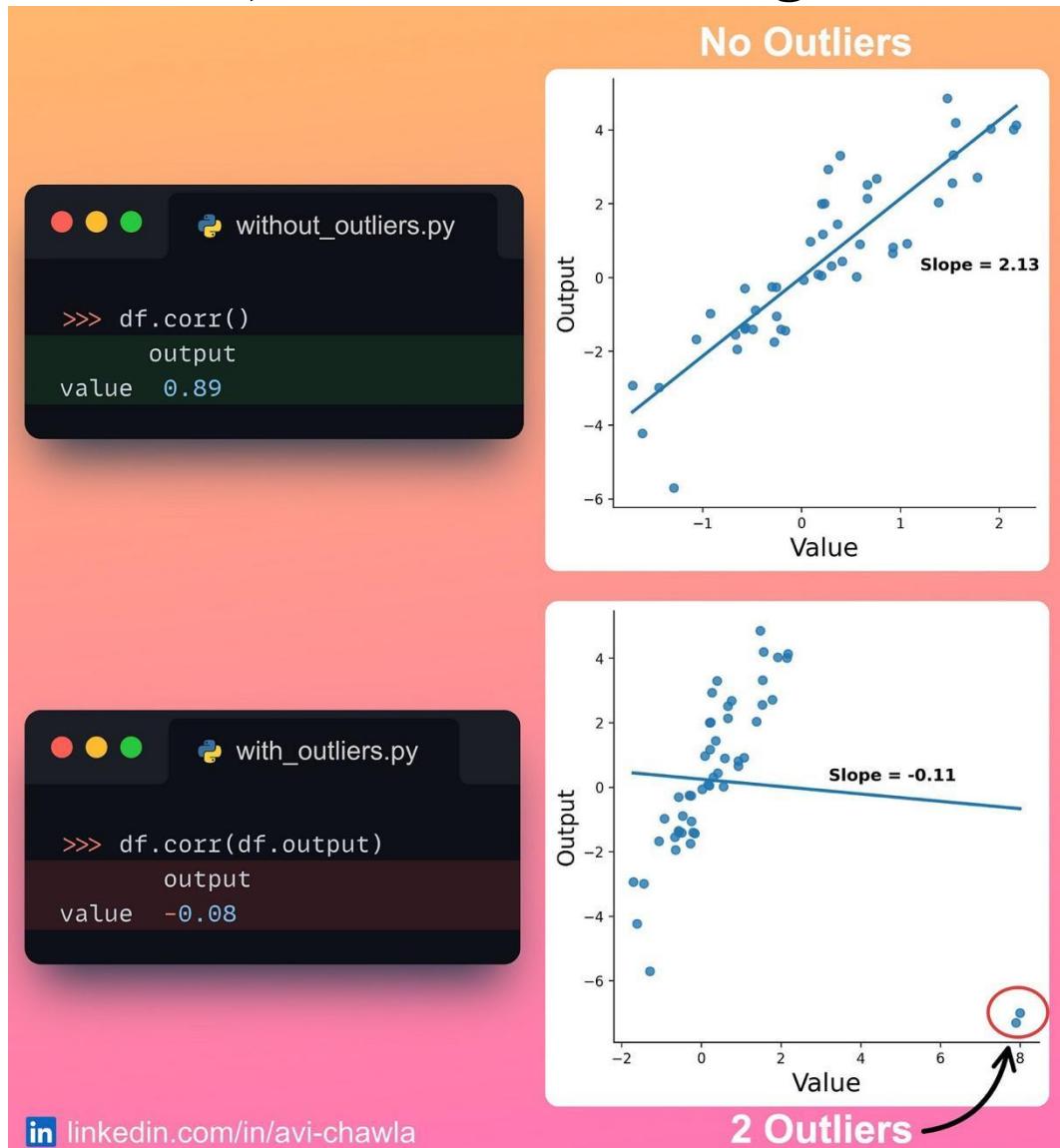
A Dot plot can be a better choice in such cases. They are like scatter plots but with one categorical and one continuous axis.

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)
```

value_counts()

City	Count	Percent	Cumulative Count	Cumulative Percent
West Jamesview	120	12.00%	120	12.00%
Aliciafort	113	11.30%	233	23.30%
New Cindychester	106	10.60%	339	33.90%
Ricardomouth	106	10.60%	445	44.50%
Whiteside	104	10.40%	549	54.90%
Kristaburgh	97	9.70%	646	64.60%
Wardfort	96	9.60%	742	74.20%
New Russellton	93	9.30%	835	83.50%
Whitakerbury	87	8.70%	922	92.20%
North Melissafurt	78	7.80%	1,000	100.00%

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

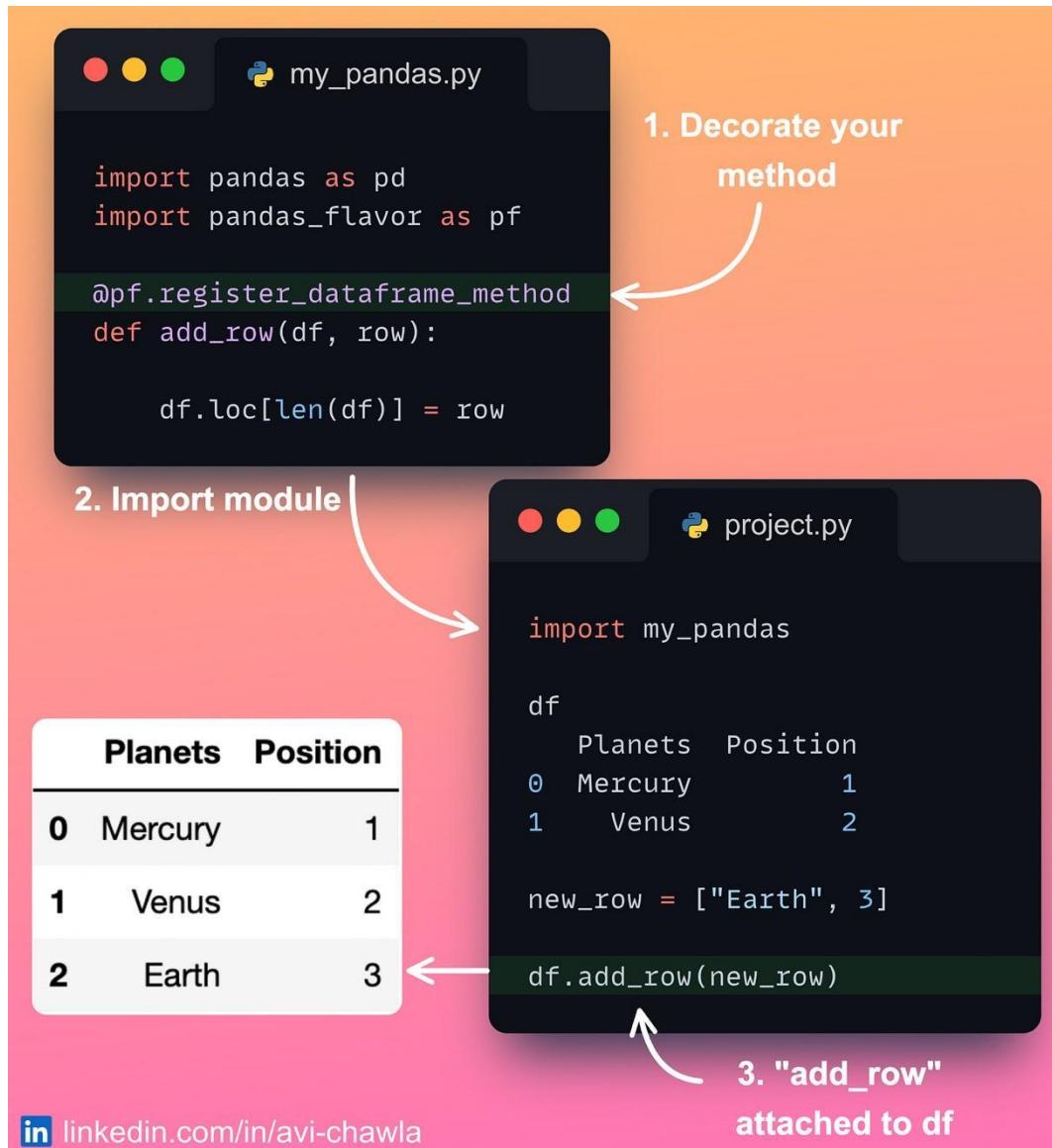
Instead, use **sidetable**. Consider it as a supercharged version of **value_counts()**. As shown below, the **freq()** method from sidetable provides a more useful summary than **value_counts()**.

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. At the top right is the logo and text "CodeSquire.ai". Below it is a code editor window containing Python code for data preprocessing:

```
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'])
```

To the right of the code editor, there are two labels: "Write comment" with a downward arrow pointing to the code, and "Output" with an upward arrow pointing from the code editor towards the generated output below.

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 Jupyter Notebook interface with two code cells and a data frame preview.

Vectorized.py:

```
df.Name.str.split()  
## 1.7 s
```

Non-vectorized.py:

```
def name_split(s):  
    return s.split()  
  
df.Name.apply(name_split)  
## 862 ms
```

df Name

	df	Name
0		Beth Alvarez
1		Deborah Watkins
2		Jeffrey Compton
3		Alan Wolfe
4		Kathryn Gordon

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

```
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
```

No type checks

No length checks

No unpacking

[in](https://linkedin.com/in/avi-chawla) 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](https://docs.python.org/3.10/tutorial/controlflow/match.html).



Enrich Your Notebook With Interactive Controls

```
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(...)
```



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

The screenshot shows a Jupyter Notebook window titled "notebook.ipynb". It contains two code cells:

```
In [1]: %load_ext jupyternotify  
In [2]: %%notify
```

Below the cells is a placeholder text: <<--YOUR-CODE-->>

A white arrow points from the bottom right of the notebook window to a browser notification. The notification is for the "Jupyter Notebook" at "localhost:8888" and says "Cell execution has finished!". It includes a Python logo icon and a "Logout" button.

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

In [1]:

```
1 import mitosheet
2 mitosheet.sheet('analysis_to_replay="id-ymyxvhaoes")
```

executed in 4.97s, finished 14:48:34 2022-11-22

The screenshot shows a Jupyter notebook cell with the code above. Below the code is a Mito spreadsheet interface displaying a pandas DataFrame. The DataFrame has columns: Name, Company_Name, Employee_City, Employee_Salary, Employment_Status, and Employee_Rating. The data consists of 100 rows of employee information. At the bottom of the Mito interface, there are buttons for '+ employee_dataset' and 'graph0'. A status bar at the bottom right indicates '(100 rows, 6 cols)'.

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

The diagram illustrates the refactoring of a Python script from using multiple if-else conditions to using a dictionary. On the left, a code editor window titled "if_else.py" contains the following code:number = int(input())
if number == 1:
 func1()

elif number == 2:
 func2()

else:
 func3()A white arrow points from the "if-else" code to a second code editor window on the right, titled "dict.py". This window contains the following code:number = int(input())
func_map = {1:func1,
 2:func2}

func_map.get(number, func3)()An upward-pointing arrow originates from the "func_map.get" line and points to the word "Default" below it. A text annotation "replace with dictionary" is placed next to the arrow.

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.

But with a dictionary, you can directly retrieve the corresponding function by providing it with the key. This makes your code concise and elegant.



DailyDoseofDS.com



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 code uses the `clear_output(wait=True)` method to clear the output before printing the next line. The output panel shows the final result:

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

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 image shows two Jupyter Notebook cells and their corresponding output tables.

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

[in 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



```
Without_slots.py
```

```
class Person:  
    def __init__(self, name, age):  
        self.Name = name  
        self.Age = age  
  
person = Person('Mike', 22)  
  
person.name = 'Peter'  
## No Error
```

'Name' mistakenly written as 'name' raises no error

```
With_slots.py
```

```
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'
```

defining __slots__ raises AttributeError

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

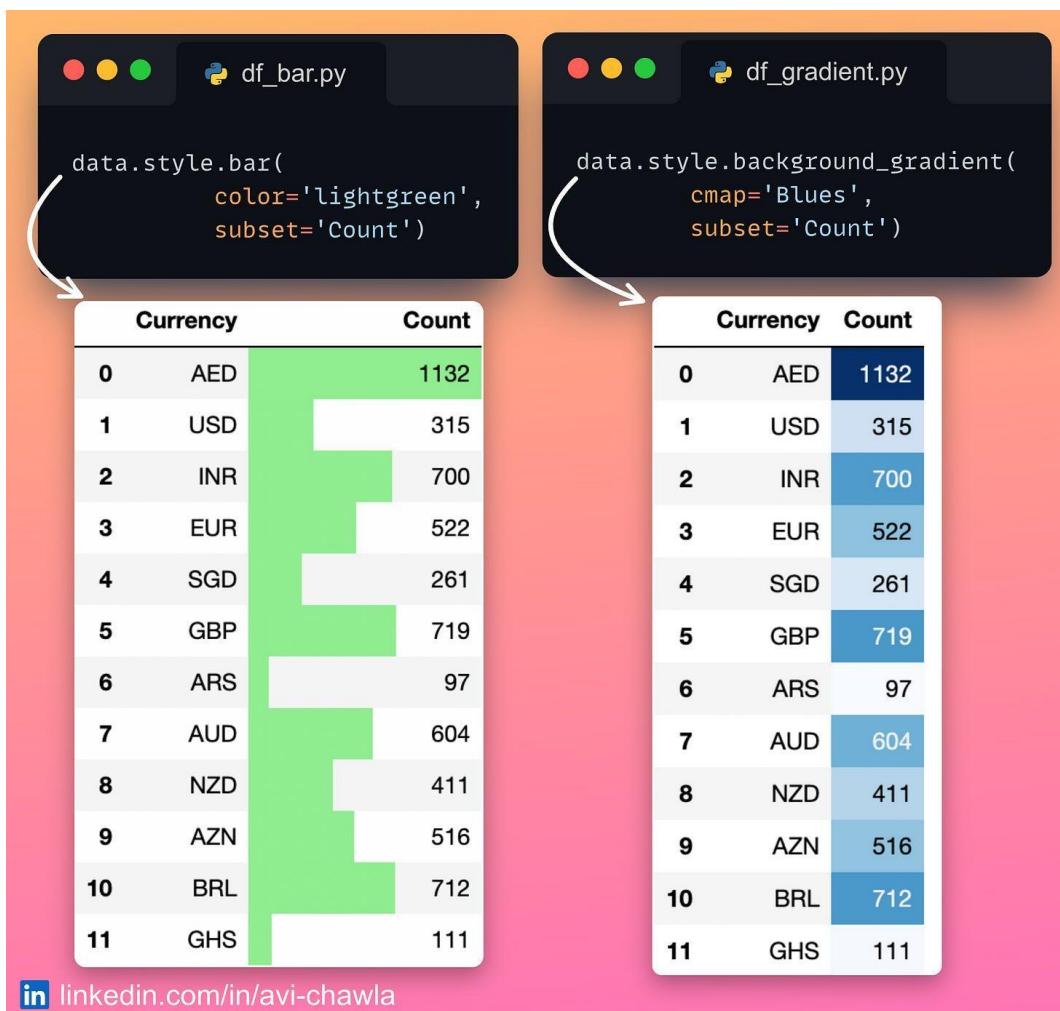
If you want to fix the attributes a class can hold, consider defining it as a slotted class.

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

data.csv

Name	Marks	Grade
Joe	95	A
Hanna	89	B
Chris	92	A
Julie	94	A

Excel to CSV

```
$ in2csv data.xlsx > data.csv
```

Column Names

```
$ csvcut -n data.csv
1: Name
2: Marks
3: Grade
```

Column Stats

```
$ csvstat data.csv
2. "Marks"
Type of data: Number
Contains null values: False
Unique values: 4
Smallest value: 89
Largest value: 95
Sum: 370
Mean: 92.5
Median: 93
StDev: 2.646
```

Query

```
$ csvsql --query "select * from data where Marks>90" data.csv
| Name | Marks | Grade |
| ----- | ----- | ----- |
| Joe | 95 | A |
| Chris | 92 | A |
| Julie | 94 | A |
```

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

A screenshot of a Mac OS X terminal window titled "fake_data.py". The window shows Python code demonstrating the Faker module. The code imports Faker, creates a Faker instance, and then uses various methods to generate fake data: name ('Darrell Alexander'), email ('ryanrichard@example.com'), address ('205 Brown Point, West Melissaport, MN 93828'), company ('Lam, Thomas and Cooper'), date_of_birth (datetime.date(1973, 1, 21)), and color_name ('LightBlue').

```
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'
```

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.

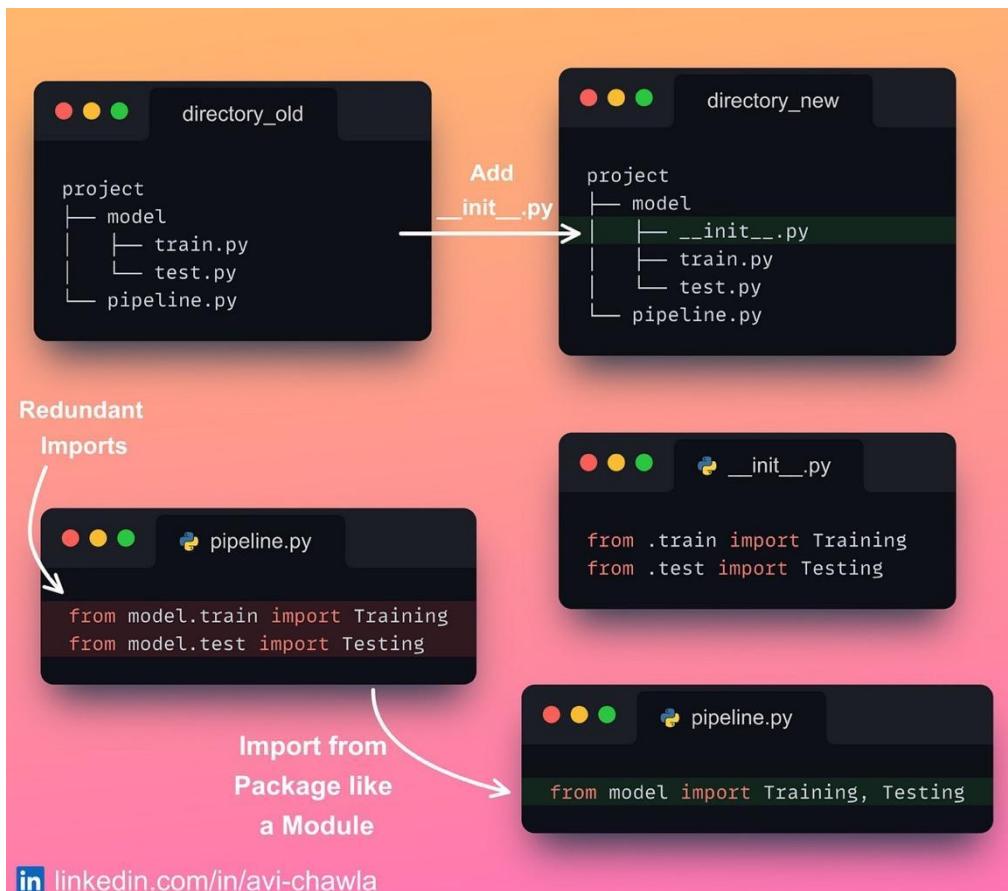
Here, looking for open-source datasets can get time-consuming. Moreover, it's possible that the dataset you find does not fit pretty well into your requirements.

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

```
number of loops (1000)           number of runs (4)  
notebook.ipynb  
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)
```

in 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



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

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.

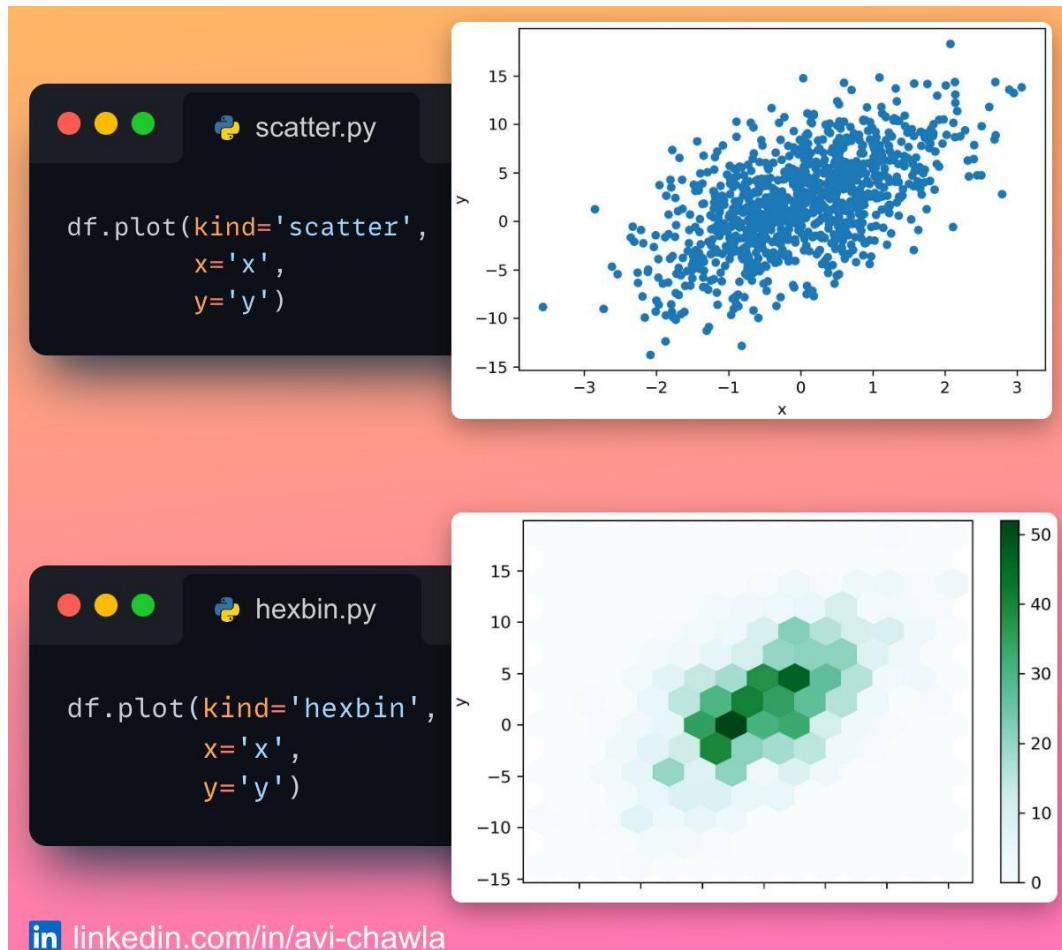
Instead, you can use a waterfall chart, which elegantly depicts these rolling differences.

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 diagram illustrates the convenience of Pyforest. On the left, a terminal window shows the standard way of importing multiple Python libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sys

from sklearn.linear_model
import LinearRegression
```

A white curved arrow points from this code to another terminal window on the right, which shows the Pyforest import statement:

```
from pyforest import *
```

Below this, four specific imports are listed, each followed by a green checkmark indicating they are successfully imported:

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.



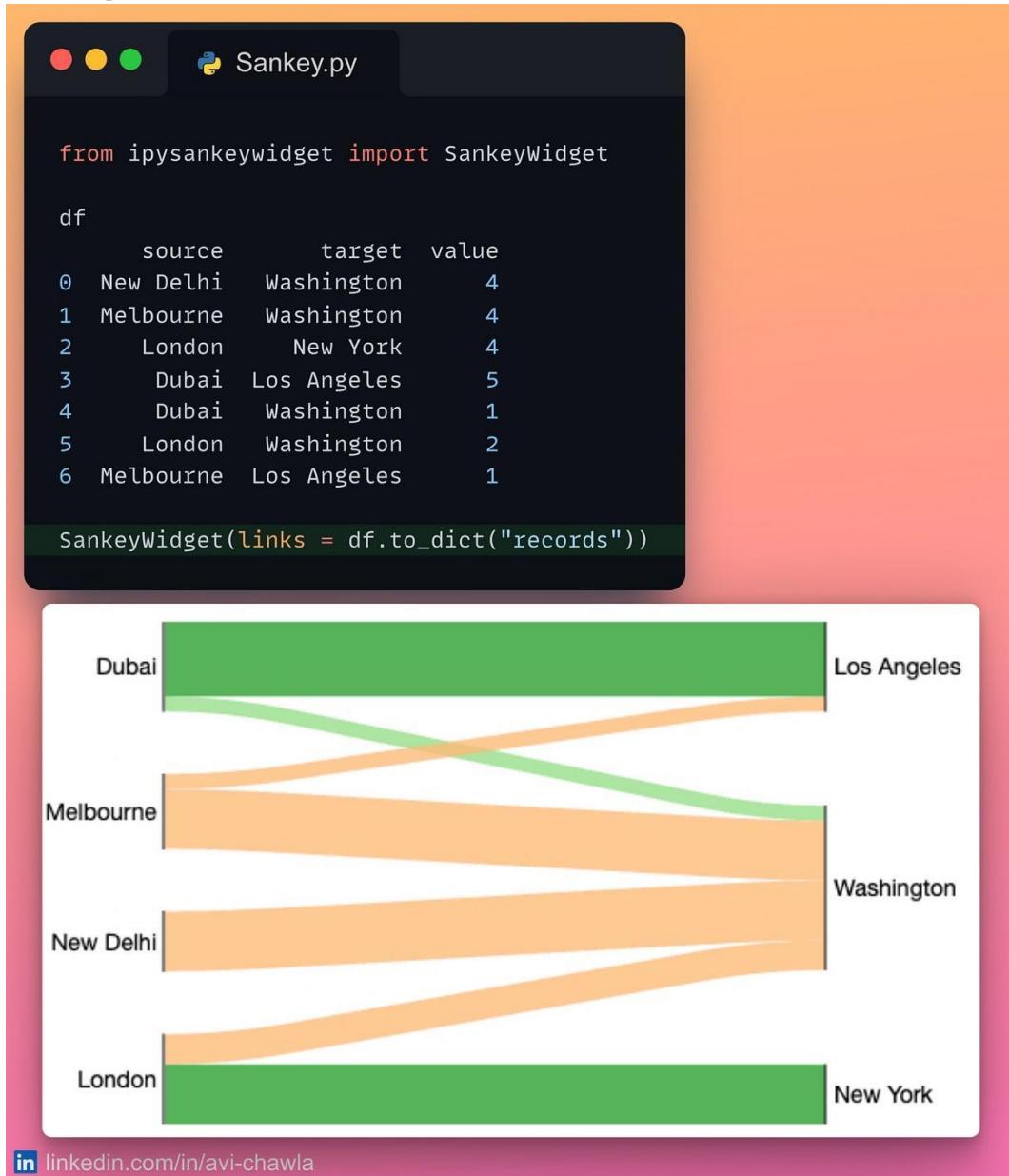
With that, you should also note that it is a good practice to keep Pyforest limited to prototyping stages. This is because once you say, develop and open-source your pipeline, other users may face some difficulties understanding it.

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

```
from sklearn.preprocessing
import PolynomialFeatures

df
  col_A  col_B      array([[ 1.,  1.,  2.,  1.,  2.,  4.],
  0      1      2          [ 1.,  3.,  4.,  9., 12., 16.],
  1      3      4          [ 1.,  5.,  6., 25., 30., 36.]])
  2      5      6
```

PolynomialFeatures().fit_transform(df)

```
from sklearn import set_config

set_config(transform_output = "pandas")

df
  col_A  col_B      1  col_A  col_B  col_A^2  col_A*col_B  col_B^2
  0      1      2      0  1.0  1.0  2.0  1.0  2.0  4.0
  1      3      4      1  1.0  3.0  4.0  9.0  12.0 16.0
  2      5      6      2  1.0  5.0  6.0 25.0  30.0 36.0
```

PolynomialFeatures().fit_transform(df)

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

The image shows a Python code editor with two code snippets. The top snippet demonstrates the use of f-strings without explicitly writing the variable names:

```
Count = 2
Fruit = "Apple"

print(f"Count = {Count}")
print(f"Fruit = {Fruit}")

## Count = 2
## Fruit = Apple
```

An annotation on the right side of this snippet reads: "Don't write variable name explicitly". A white arrow points from this annotation to the first line of the code.

The bottom snippet shows the same code after adding an equals sign (=) inside the curly braces to print the variable names along with their values:

```
print(f"{Count = }")
print(f"{Fruit = }")

## Count = 2
## Fruit = Apple
```

An annotation on the left side of this snippet reads: "Add '=' in curly braces {}". A white arrow points from this annotation to the first line of the code.

At the bottom left of the code editor, there is a LinkedIn icon followed by the URL: 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 side-by-side. The left window has a yellow-to-orange gradient background and contains the following Python code:

```
import time

start = time.time()
time.sleep(10)
end = time.time()

print(end - start)
## 10.00482
```

The right window has a pink-to-red gradient background 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 terminals is a blue LinkedIn profile link: 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](https://www.linkedin.com/in/avi-chawla)

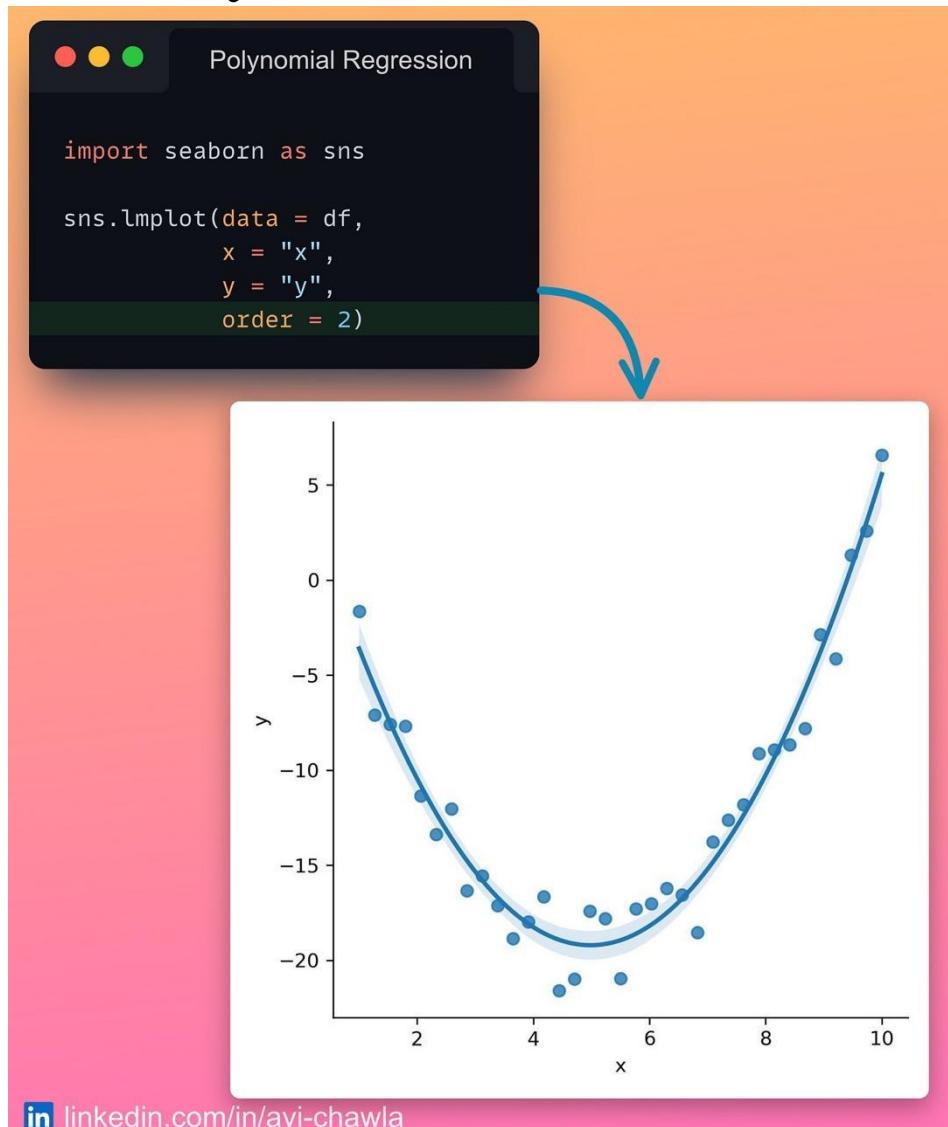
DALL·E is now accessible using the OpenAI API.

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



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

The screenshot shows a Jupyter Notebook interface with two code cells and their outputs.

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

An orange arrow points from the "Out[3]" output cell to the "Out[4]" input cell.

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://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 image shows a Jupyter Notebook interface with two code cells and their outputs.

Step 1: Copy Table

```
# b    too    one   b  12
# 7    foo    three 7  14

print(df.loc[df['A'] == 'foo'])
```

yields

	A	B	C	D
0	foo	one	0	0
2	foo	two	2	4
4	foo	two	4	8
6	foo	one	6	12
7	foo	three	7	14

Step 2: Read in Pandas

```
import pandas as pd

df = pd.read_clipboard()
```

```
>>> df.head()
      A      B  C  D
0  foo    one  0  0
2  foo    two  2  4
4  foo    two  4  8
6  foo    one  6 12
7  foo  three  7 14
```

[in.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.

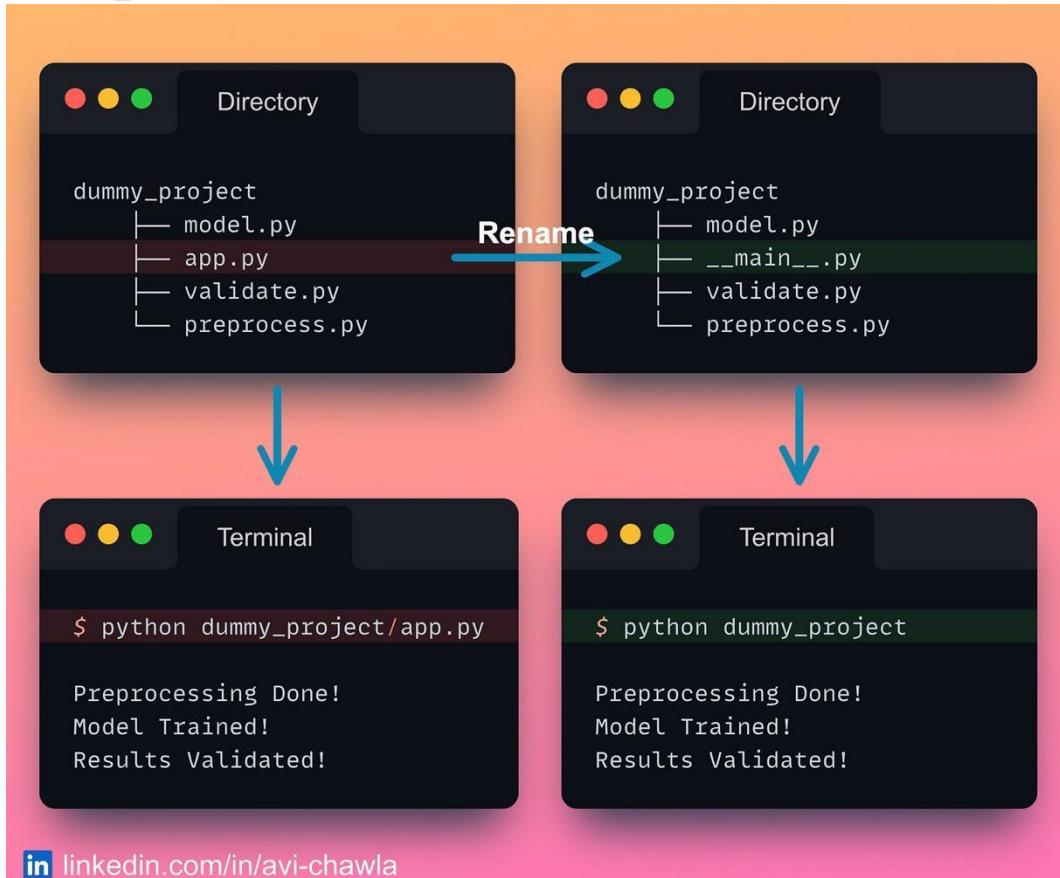
Instead, with the **read_clipboard()** method, you can eliminate the CSV step altogether.

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

```
file.py
```

```
1 def func():
2     print(0)
3     ...
4
5     if condition:
6         print(1)
7         ...
8     else:
9         print(2)
10    ...
```

```
Terminal
```

```
$ python file.py
0
2
```



```
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
```

```
$ 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

```
● ○ ● 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
```

```
● ○ ● Numpy Where  
  
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

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.

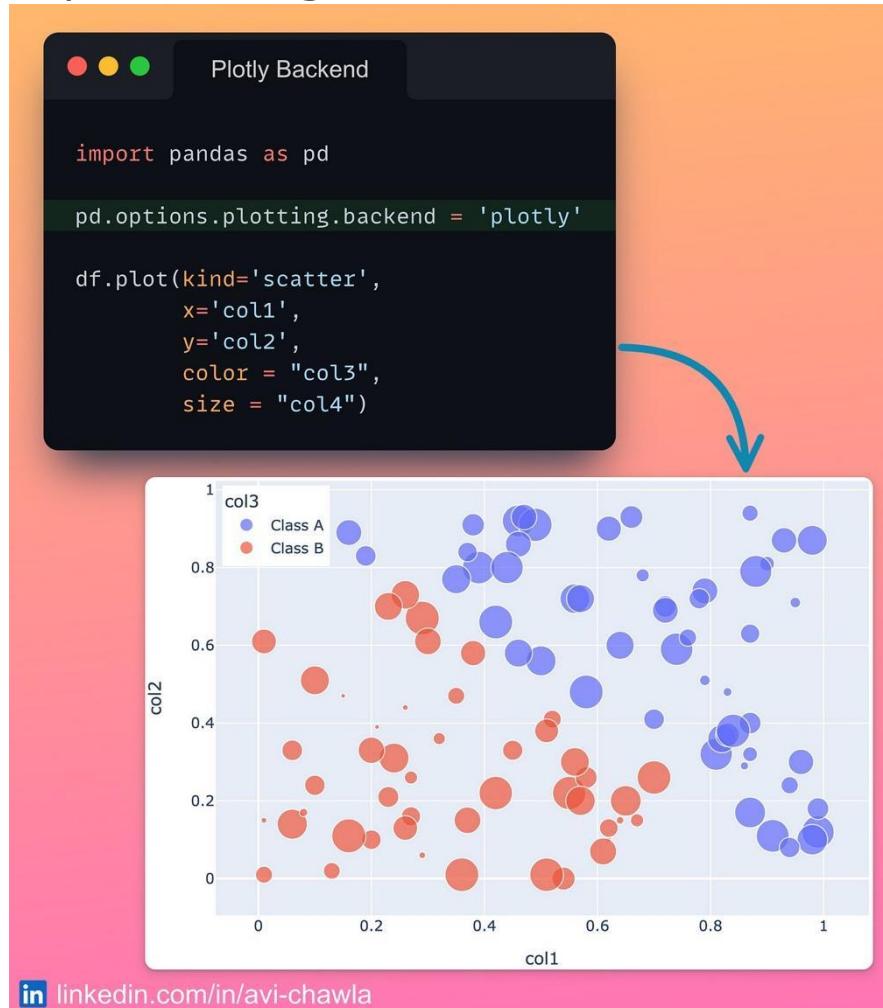
Instead, you should use the **np.where()** method to create conditional columns. It does the same job but is extremely fast.

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

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



Search keyword

Search in All Notebooks

Command Line

```
$ grep "polyfit" *.ipynb
```

```
numpy_lr.ipynb: "z = np.polyfit(x, y, deg = 2)"  
numpy_lr.ipynb: "z = np.polyfit(x, y, deg = 1)"
```

linkedin.com/in/avi-chawla

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

The diagram illustrates a workflow for persisting variables across kernel restarts in a Jupyter Notebook.

Initial State: A Jupyter Notebook window titled "Notebook.ipynb" shows two cells:

- [1]: `value = 10`
- [2]: `%store value`

After Restart: A second Jupyter Notebook window titled "Notebook.ipynb" shows the results of the stored variable:

- [1]: `%store -r value`
- [2]: `value`
10

A large blue arrow labeled "Restart" points from the initial state to the second window, indicating the process of restarting the kernel.

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

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.

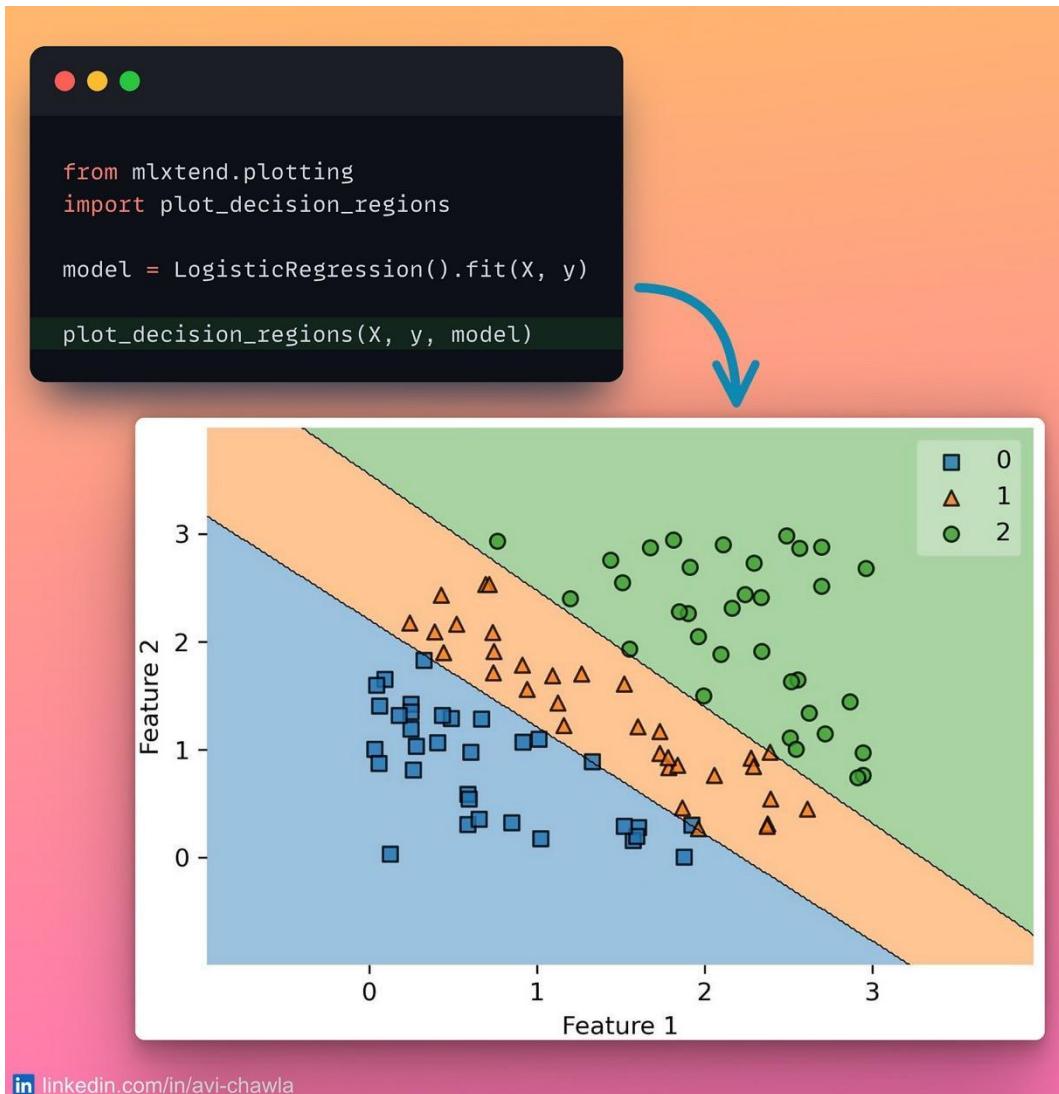


The performance gain is not just limited to I/O but is observed in many other tabular operations as well.

Read more here: [DataTable Docs](#).



Elegantly Plot the Decision Boundary of a Classifier



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 cells side-by-side. The left cell demonstrates importing multiple metrics from `sklearn.metrics` and using them individually:

```
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
```

A blue arrow points from the text "Import all metrics individually" to this cell.

The right cell demonstrates using `get_scorer` to get a scorer object from a string name:

```
from sklearn.metrics import get_scorer

accuracy = get_scorer("accuracy")
>>> accuracy._score_func(y_true, y_pred)
0.5

precision = get_scorer("precision")
>>> precision._score_func(y_true, y_pred)
0.8
```

A blue arrow points from the text "Get a scorer from string" to this cell.

Below the notebooks is a LinkedIn profile link:

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_scorer()` 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 arrow from the bottom cell points to this line, with the text 'Output is Pandas DataFrame' next to it. The bottom cell, titled 'Scikit-learn 1.2.dev', shows the same code but includes a 'set_output' step:scaler = StandardScaler()
scaler.set_output(transform="pandas")
X_scaled = scaler.fit_transform(X_train)

type(X_scaled) ## pandas.core.frame.DataFrameTo the right of this cell, the text 'Output is NumPy Array' is displayed.

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

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)
```

100% [██████████] 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 image shows two screenshots of a Jupyter Notebook interface. The left screenshot displays Python code that prints odd numbers from 1 to 19. The right screenshot shows the same code with an additional print statement for even numbers, demonstrating the modification of a function during run-time. A red arrow points from the left code to the right code, labeled '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:
        pass

for i in range(100):
    func(i)
```

```
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

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.com/in/avi-chawla

Have you ever been in a situation where you wished to add more details to an already running code?

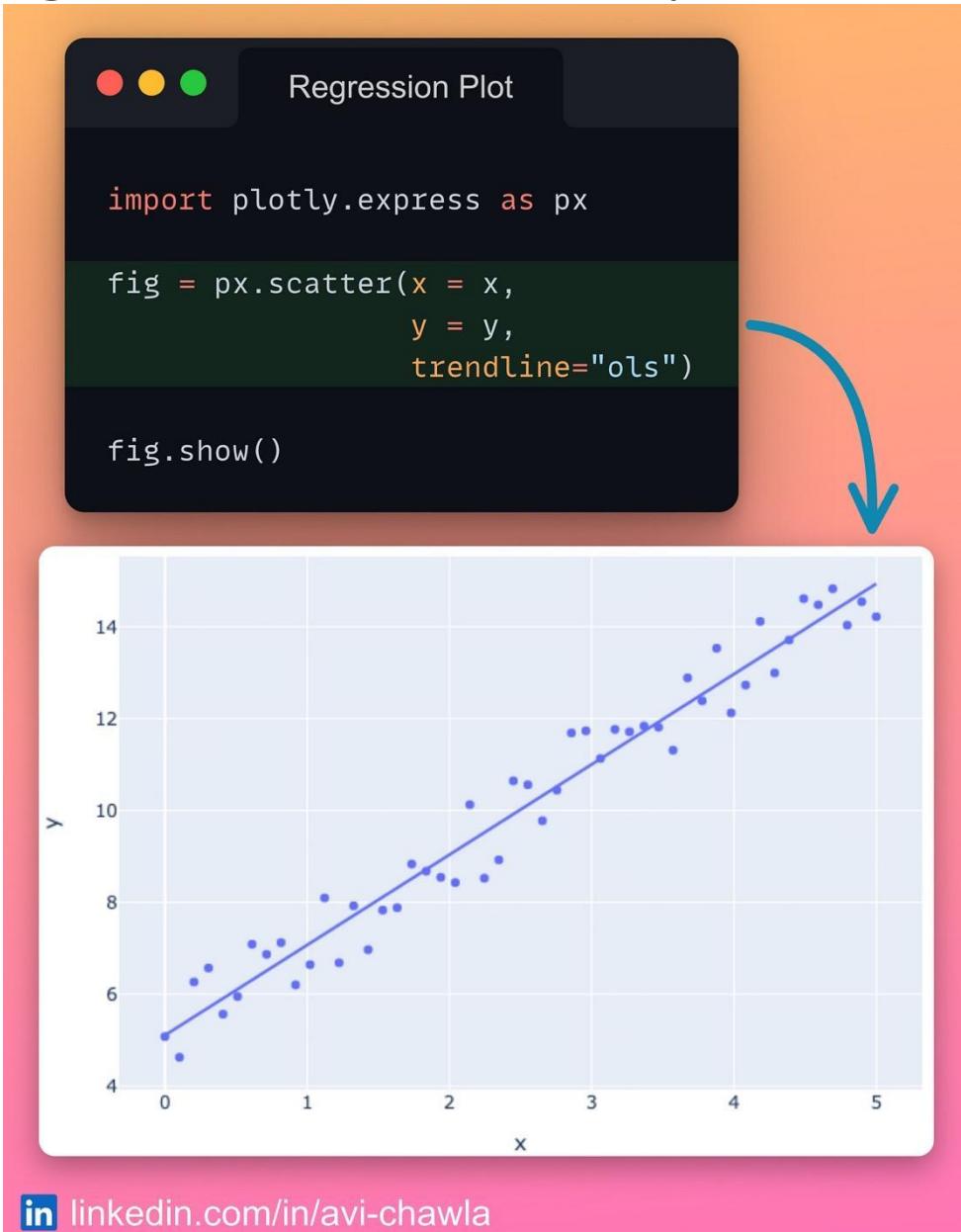
This is typically observed in ML where one often forgets to print all the essential training details/metrics. Executing the entire code again, especially when it has been up for some time is not an ideal approach here.

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



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 specific lines of code to their descriptions:

- An arrow points from the line `x = np.hstack((x, x**2))` to the text "Create Polynomial Features".
- An arrow points from the line `deg = 2` to the text "Specify Degree".

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

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 code snippets in a Jupyter Notebook environment.

Top Snippet:

```
>>> df
   col1  col2  col3  col4
0      1      7      4      A
1      3      9      6      B
2      6      2      5      A
```

Middle Snippet:

```
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 middle snippet.

Bottom Snippet:

```
df = df.astype({
    "col1":np.int32,
    "col2":np.int16,
    "col3":np.float16})
```

A blue arrow points from the text "Single Call" to the bottom snippet.

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

```
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

```
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.

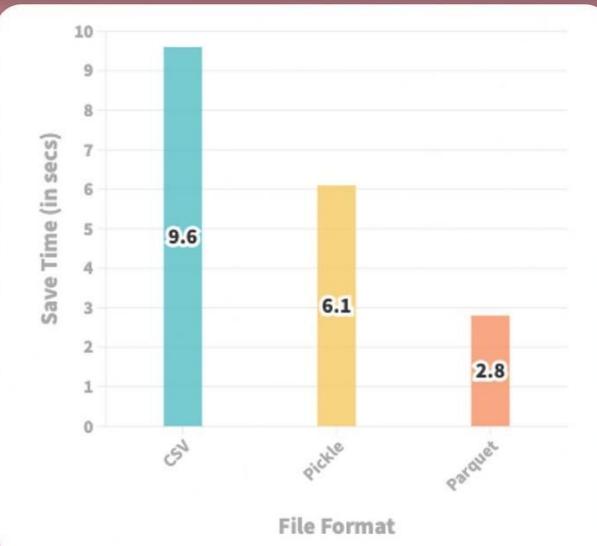
"Pandarallel" allows you to parallelize its operations to multiple CPU cores - by changing just one line of code. Supported methods include apply(), applymap(), groupby(), map() and rolling().

Read more: [GitHub](#).



Why you should not dump DataFrames to a CSV

```
Save DF  
  
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

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 terminal windows side-by-side. The left window is titled 'List.py' and the right window is titled 'Generator.py'. Both windows use a dark theme.

List.py Content:

```
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
```

Generator.py Content:

```
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
```

At the bottom left of the terminal area, there is a small LinkedIn icon followed by the URL linkedin.com/in/avi-chawla.

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
```

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

The image shows a Mac OS X desktop environment. On the left, a code editor window titled "code.py" displays the following Python code:

```
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)
```

On the right, a terminal window titled "Terminal" shows the output of the command "\$ vulture code.py". A blue arrow points from the terminal window up towards the code editor window.

```
$ 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

The slide features two code snippets side-by-side. The left snippet, titled 'Categorical Col.', shows how to check the data type of a 'Gender' column in a DataFrame:

```
import pandas as pd
len(df.Gender)
## 1500
df.Gender.unique()
## ["Male", "Female"]
```

The right snippet, titled 'Reduce Memory Usage', demonstrates converting the 'Gender' column to a categorical type and measuring the memory savings:

```
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
```

At the bottom left of the slide is a LinkedIn profile link: linkedin.com/in/avi-chawla.

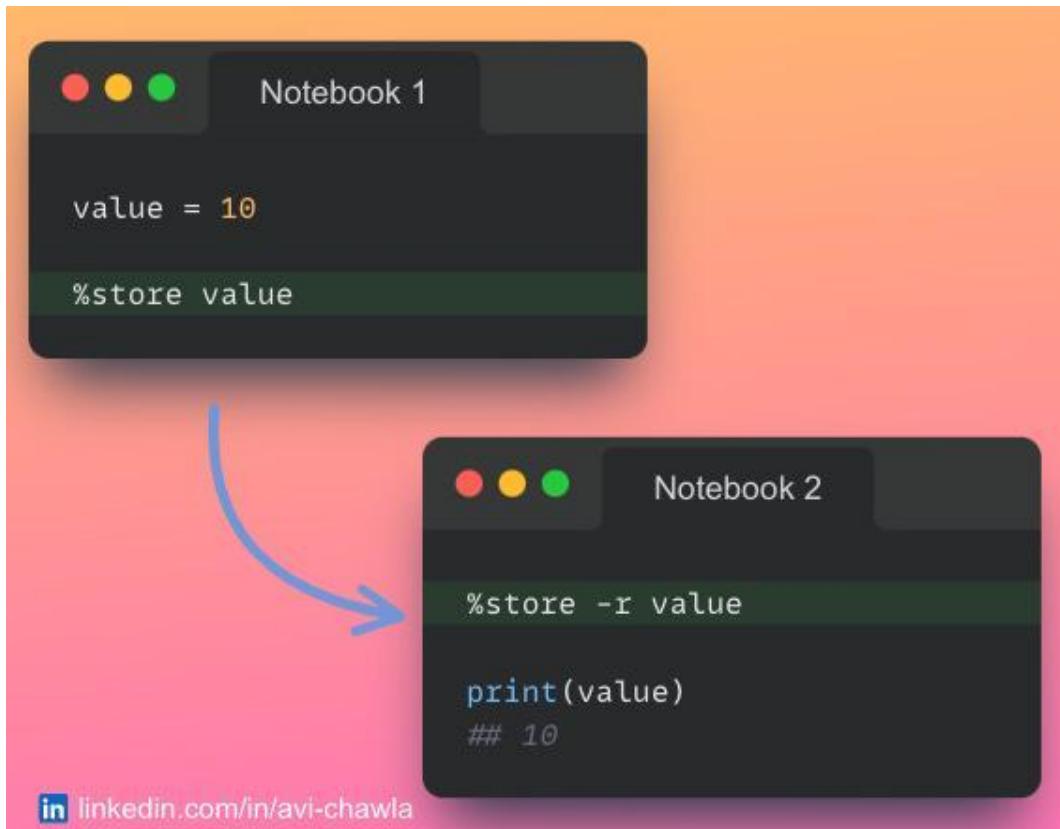
If your data has categorical columns, you should not represent them as int/string data type.

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

The screenshot shows two terminal windows side-by-side. The left window is titled 'Pandas' and the right window is titled 'Datatable'. Both windows show Python code for reading a CSV file named 'file.csv'. The Pandas code takes approximately 8.82 seconds, while the Datatable code takes approximately 4.04 seconds. The Datatable code is also more concise, using fewer lines of code.

```
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 image shows two screenshots of a Mac OS X terminal window. In the top screenshot, the code prints odd numbers from 1 to 11 as 'Odd'. In the bottom screenshot, the code has been modified to also print even numbers as 'Even'. A red arrow points from the top terminal to the bottom one, labeled 'Modified During Run-time'.

Top Terminal Output:

```
from time import sleep
from reloading import reloading

for number in reloading(range(100)):

    if number % 2:
        print(f"{number} is Odd")
    else:
        pass
```

Bottom Terminal Output:

```
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")
```

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

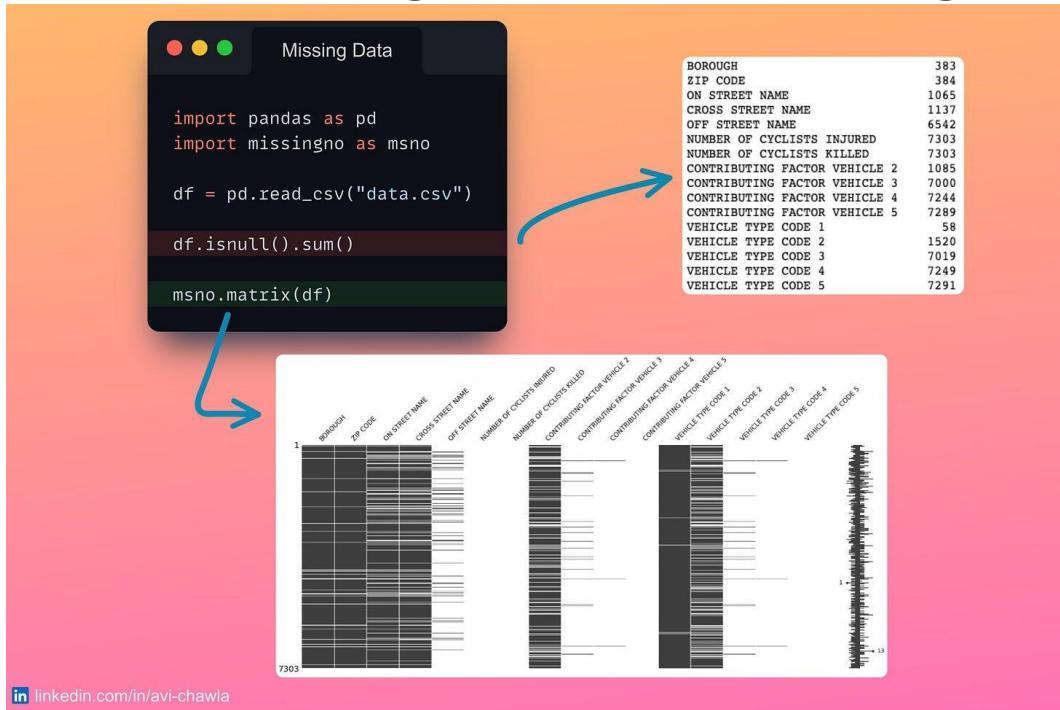
Have you ever been in a situation where you wished to add more details to an already running code (printing more details in a for-loop, for instance)?

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.