


Nine Emerging Python Libraries You Should Add to Your Data Science Toolkit in 2021

 towardsdatascience.com/nine-emerging-python-libraries-you-should-add-to-your-data-science-toolkit-in-2021-85ce3d239846

September 29, 2021



Photo by on

If you enjoy this, be sure to to never miss another article on data science guides, tricks and tips, life lessons, and more!

As Data Science continues to grow and develop, it's only natural for new tools to emerge, especially considering the fact that data science had some significant barriers to entry in the past.

In this article, I wanted to go over nine libraries that I've come across in the past year that are **game changers**. These libraries have been incredibly useful in my data science journey and I wanted to share them with you in hopes that it'll help you with your journey too!

The following libraries are broken down into three categories:

1. Model Deployment
2. Data Modelling
3. Exploratory Data Analysis

1. Model Deployment

Kedro

It's no surprise that data science continues to converge more and more with software engineering practices, given that data science is extremely computer science dependent. As data science continues to evolve, a number of solutions are being made to help make it easier to productionalize data science solutions.

One of these solutions includes Kedro.

Kedro is a workflow tool for data science pipeline development that encourages production-ready codes and allows you to build portable pipelines for your data. Overall, it applies software engineering principles to help you make your code more standardized, reproducible, and modular.

A new Python library for production-ready data pipelines

Kedro is an open-source Python framework for creating reproducible, maintainable and modular data science code. It...

Gradio

Gradio lets you build and deploy web apps for your machine learning models in as little as three lines of code. It serves the same purpose as Streamlit or Flask, but I found it much faster and easier to get a model deployed.

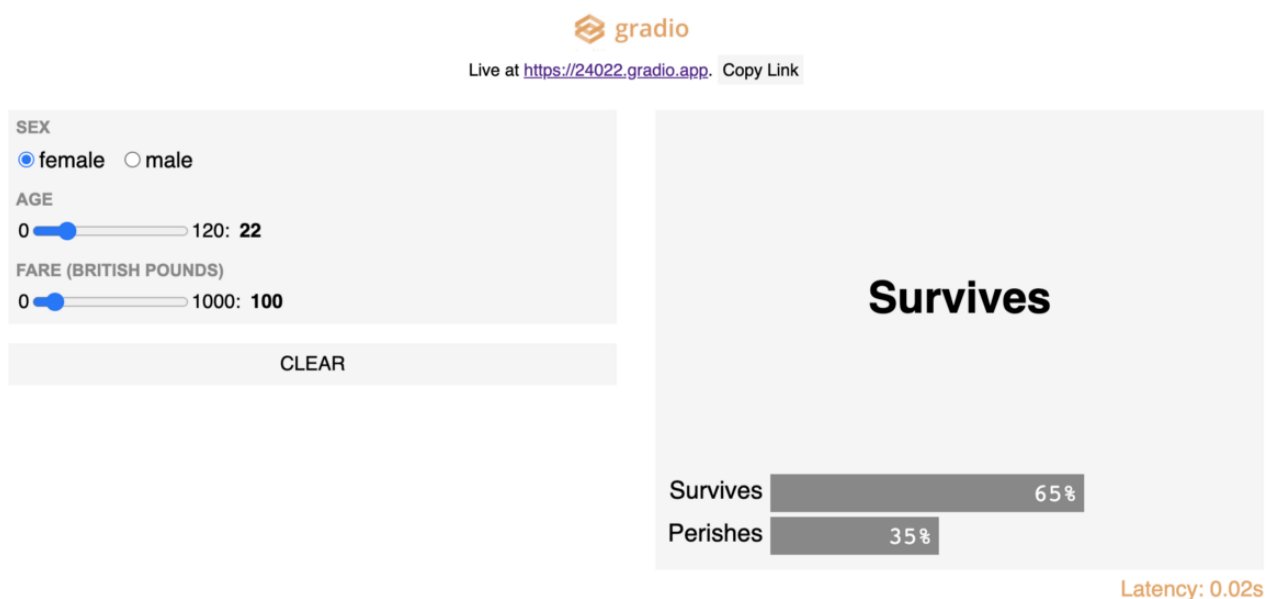


Image taken by Gradio with permission

Gradio is useful for the following reasons:

1. It allows for further model validation. Specifically, it allows you to interactively test different inputs into the model.
2. It's a good way to conduct demos.
3. It's easy to implement and distribute because the web app is accessible by anyone through a public link.

Quickly create customizable UI components around your models. Gradio makes it easy for you to "play around" with your...

Comparing several web UI tools for data science!

Streamlit

Building machine learning and data science applications and programs can be a difficult and often over-complicated process.

Streamlit is another popular tool that is used to create user interfaces. It is an open-source Python library that is used to build powerful, custom web applications for data science and machine learning. Streamlit is compatible with several major libraries and frameworks such as Latex, OpenCV, Vega-Lite, seaborn, PyTorch, NumPy, Altair, and more.

Deploying your Streamlit application to Heroku to showcase your Data Solution

The easiest way to learn how to use Streamlit is to try things out yourself. As you read through this guide, test each...

2. Data Modelling

PyCaret

There are a lot of tasks in the machine learning side of data science that we want to do quickly and get answers immediately, but can't if lengthy code bogs us down.

PyCaret is a **low-code** machine learning library that allows you to jump straight from idea to answer by creating models very quickly. This also means that you can conduct experiments, impute missing values, encode categorical data, and feature engineer much quicker than you would traditionally.

PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It is...

Train, visualize, evaluate, interpret, and deploy models with minimal code.

Prophet

Time series is a crucial concept in data science and is used every day to make helpful forecasts about a range of scenarios, for example, a retail store's revenue, or a city's crime rates. Prophet is a library for Python that allows you to create time series models and apply your data to them to get automatically updating forecasts.

Prophet has been developed by Facebook and is an extremely powerful tool, particularly for time series analyses.

Prophet is a forecasting procedure implemented in R and Python. It is fast and provides completely automated forecasts...

Time series forecasting can be challenging as there are many different methods you could use and many different...

3. Exploratory Data Analysis

Pandas Profiling

Pandas Profiling is a Python library that completes your standard EDA in one line of code. It essentially computes several analyses and displays them in the form of a report which shows you things like the characteristics of the dataset, variable properties, correlation of variables, missing values, the distribution of data, and more.

It's as simple as implementing the following:

```
from pandas_profiling import ProfileReport
profile = ProfileReport(df, title="Pandas Profiling Report")
profile
```

Documentation | Slack | Stack Overflow Generates profile reports from a pandas DataFrame. The pandas df.describe()...

Pandas Profiling, the perfect tool for exploratory data analysis.

D-Tale

If you're an Excel wizard then you'll love D-Tale.

D-Tale is a Python library that visualizes a Pandas DataFrame, but more specifically, it visualizes it in the form of a highly interactive pivot table!

D-Tale's selling point is its many features similar to Pandas Profiling, but also provides features related to excel pivot tables, like conditional formatting, sorting data, filtering data, etc.

Introduction to D-Tale for interactive data exploration

Autoviz

If Pandas Profiling and D-Tale aren't enough to automate your EDA and visualizations, then Autoviz (automated visualizations) is as good as it gets. Just like its name suggests, Autoviz turns your data into stunning demonstrations with very few lines of code.

Autoviz is fast to find important features in your data and spread it all out in front of you just from a single line of code. This makes it easy to work on large datasets and understand what's going on so you can make faster changes while being impressed by the cleanliness of your data.

AutoViz is a one-click visualization engine: You just need to upload your data in the box above and tell us your target...

autoviz.io

Visualize dataset of any size with just one line of code

Plotly

It goes without saying that graphs and demonstrations are an integral part of data science. Not only do graphs allow you to immediately see when you've broken something, but they also give you a strong visual sense of what impact your code changes make on your data.

Plotly is definitely a must-know tool for building visualizations since it is extremely powerful, easy to use, and has a big benefit of being able to interact with the visualizations.

Along with Plotly is Dash, which is a tool that allows you to build dynamic dashboards using Plotly visualizations. Dash is a web-based python interface that removes the need for JavaScript in these types of analytical web applications and allows you to run these plots online and offline.

A brief introduction to plotly.py library

How to make great-looking, fully-interactive plots with a single line of Python
