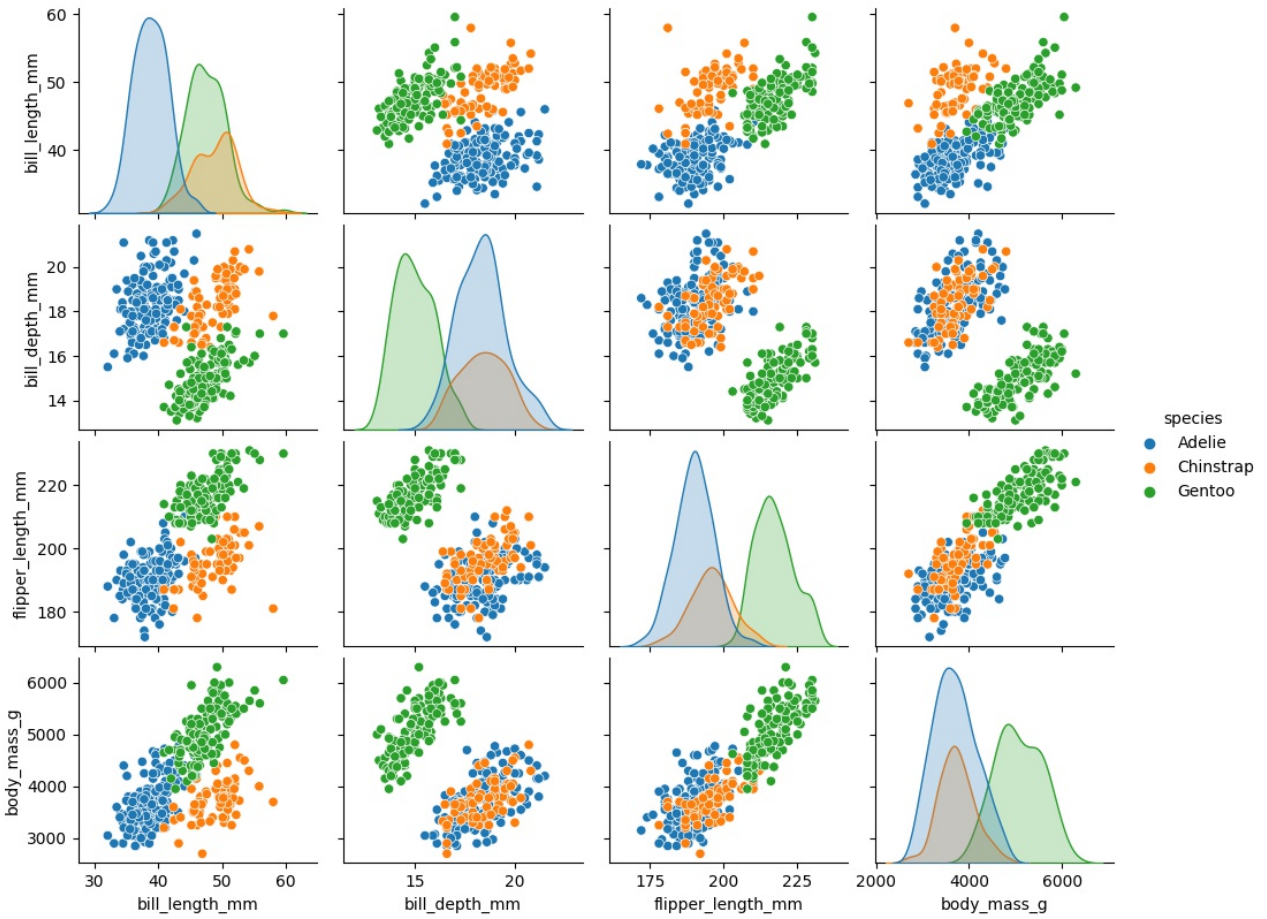
Version3
Status!
Author

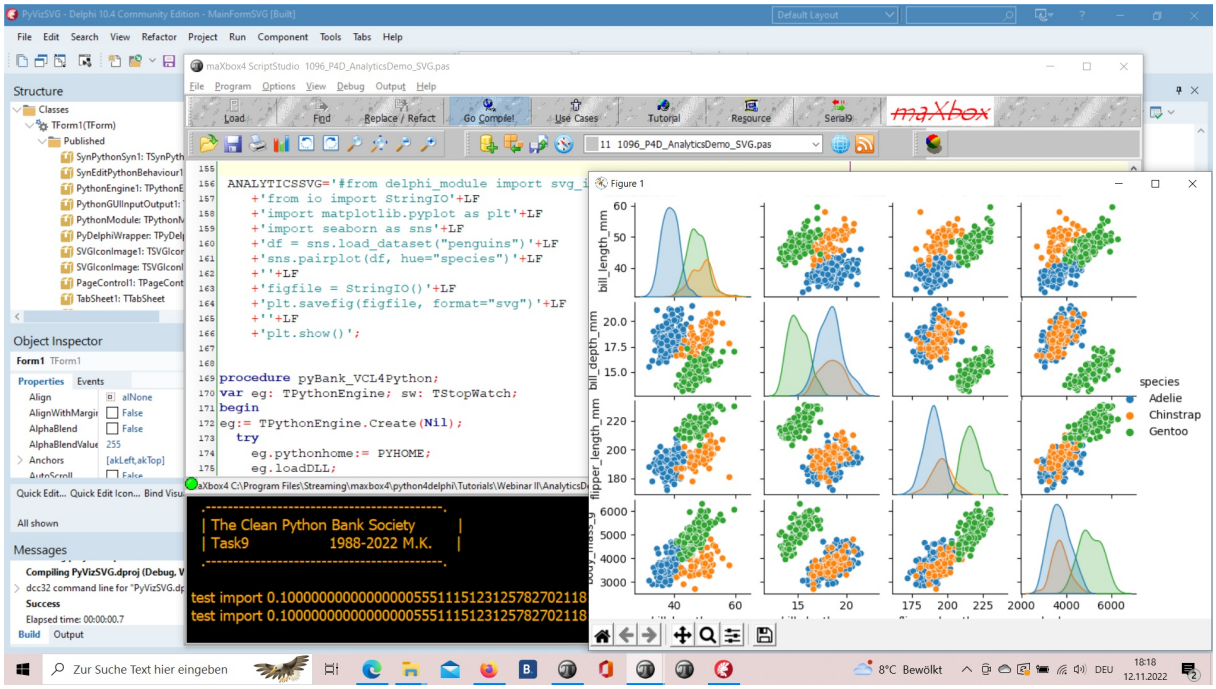
# Data Science Story

## maXbox Starter 100



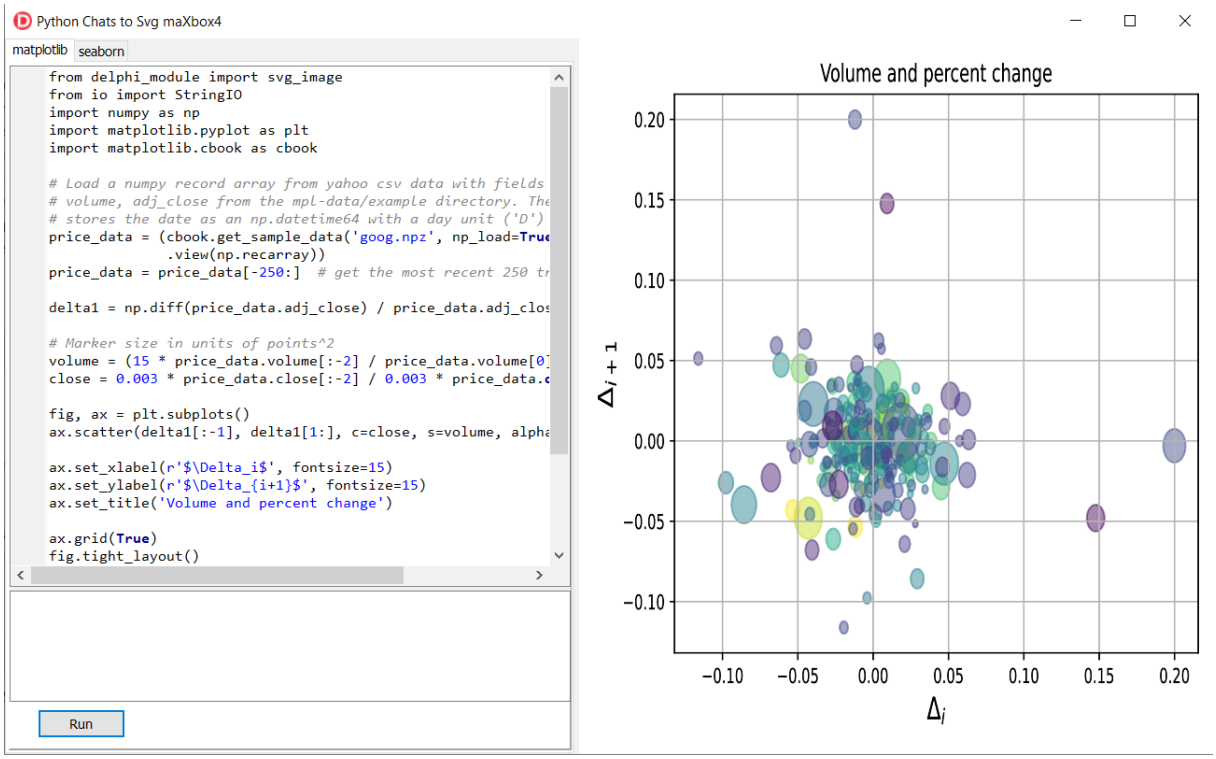
| Links and Sources |
| --- |
| Title |
| https://www.kaggle.com/code/parulpandey/penguin-dataset-the-new-iris |
| |
| python4delphi/Tutorials/Webinar II at master · maxkleiner/python4delphi (github.com) |
| |

https://maxbox4.wordpress.com/




1096_2022-11-12_svg_delphi_maxbox_seaborn4.png

RAD Studio 11.4, maXbox4 and Python 3.8 with Seaborn




Python4Delphi with SVG Plot

# 1 From Data to Plot

## 1.1 Functions

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper.

Palmer Archipelago (Antarctica) penguin dataset appears to be a drop in replacemnt for the same. It is a great intro dataset for data exploration & visualization. But the penguins dataset has different number of samples for each species.

First we get the data:

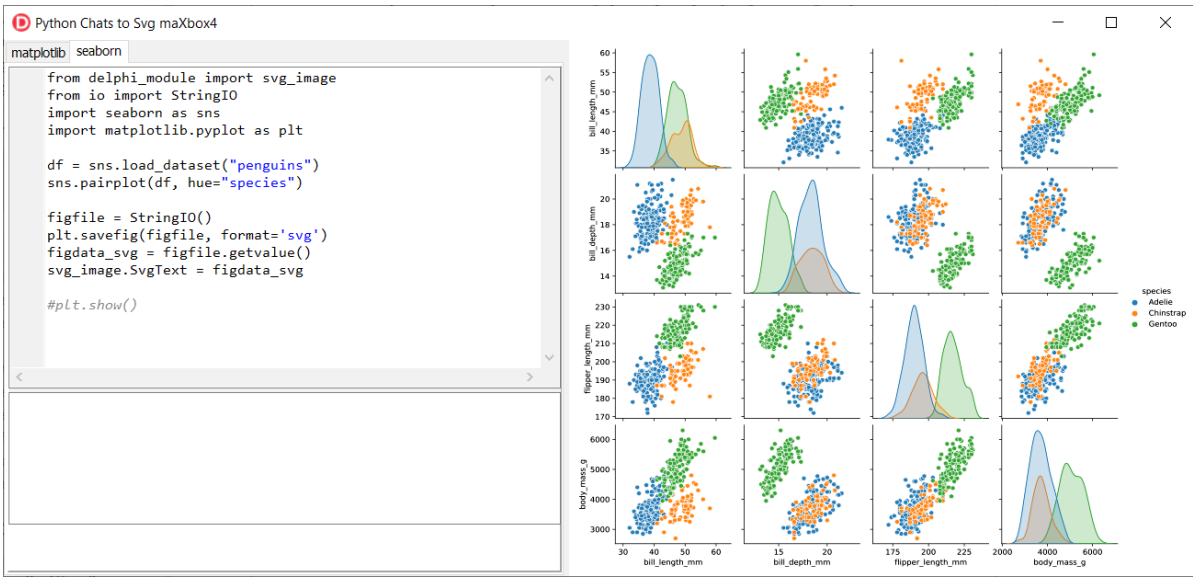
```
import seaborn as sns
df = sns.load_dataset("penguins")
```

Or alternate

```
df = pd.read_csv('../input/palmer-archipelago-antarctica-penguin-data/penguins_size.csv')
```

```
df.head()
```

| | species | island | culmen_length_mm | culmen_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|---|
| **0** | Adelie | Torgersen | 39.1 | 18.7 | 181.0 | 3750.0 | MALE |
| **1** | Adelie | Torgersen | 39.5 | 17.4 | 186.0 | 3800.0 | FEMALE |
| **2** | Adelie | Torgersen | 40.3 | 18.0 | 195.0 | 3250.0 | FEMALE |
| **3** | Adelie | Torgersen | NaN | NaN | NaN | NaN | NaN |
| **4** | Adelie | Torgersen | 36.7 | 19.3 | 193.0 | 3450.0 | FEMALE |




**1096_2022-11-12_svg_python2seaborn.png**

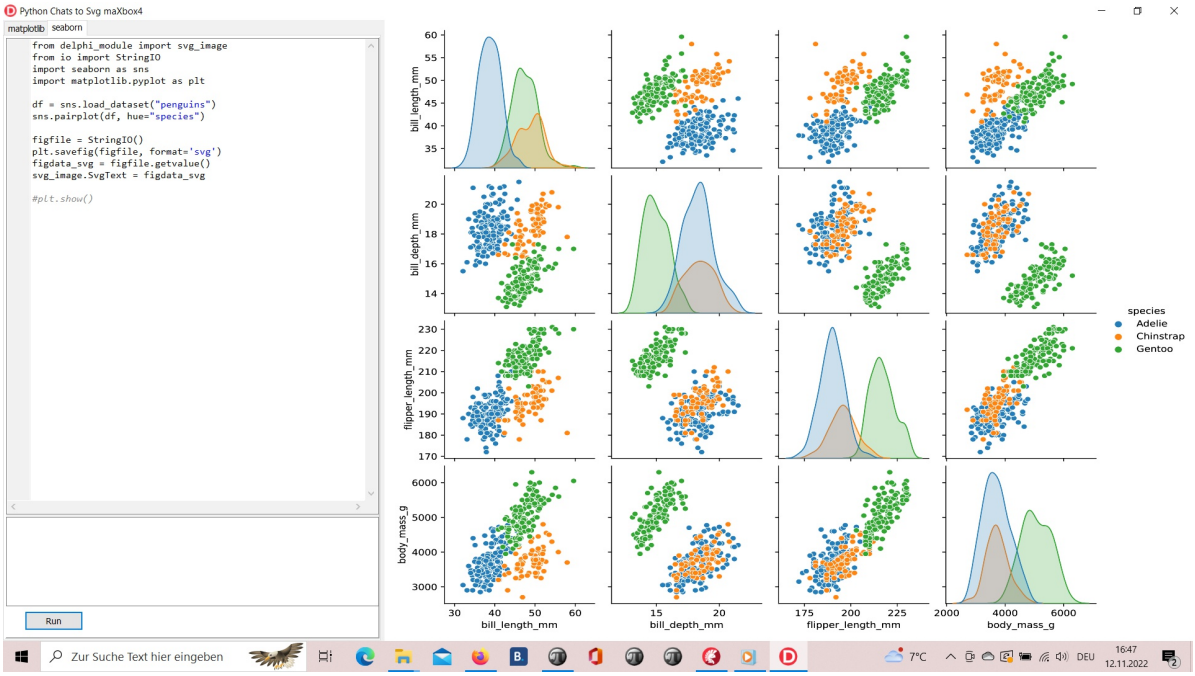
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   species            344 non-null    object
 1   island             344 non-null    object
 2   culmen_length_mm   342 non-null    float64
 3   culmen_depth_mm    342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                334 non-null    object
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

```
memory shape core cube #7
```

**The dataset consists of 7 columns.**

- **species: penguin species (Chinstrap, Adélie, or Gentoo)**
- **culmen_length_mm: culmen length (mm)**
- **culmen_depth_mm: culmen depth (mm)**
- **flipper_length_mm: flipper length (mm)**
- **body_mass_g: body mass (g)**
- **island: island name (Dream, Torgersen, or Biscoe) in the Palmer Archipelago (Antarctica)**
- **sex: penguin sex**



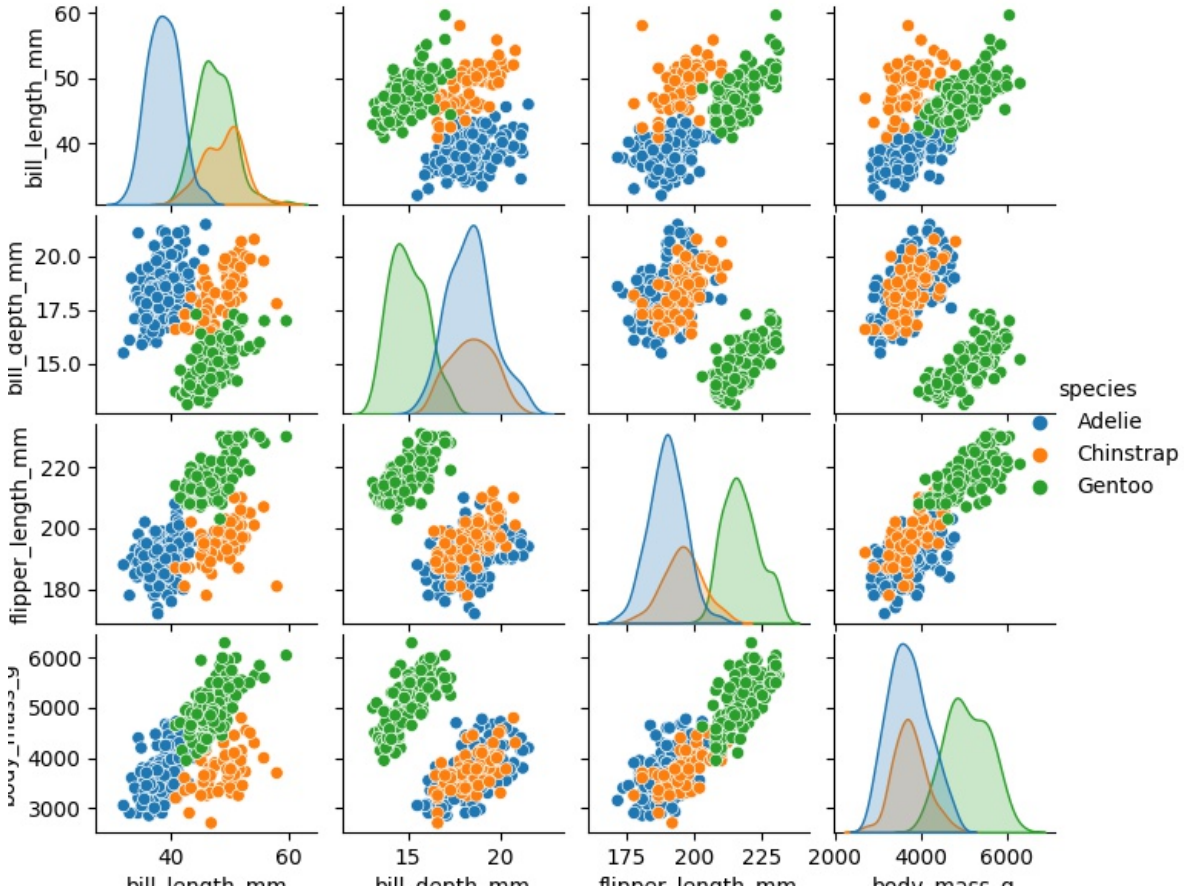

**1096_2022-11-12_svg_python_delphi_scalable.png**

**Seaborn** is a **Python** data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

## 1.2 Covariance

"Covariance" indicates the direction of the linear relationship between variables. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.

Source: https://tinyurl.com/yd2pezss




**1096_Figure_1_species.png**

The scatter plot which shows us the correlation with respect to other features. This method helps just to figure out the important features which account the most for the classification in our model.

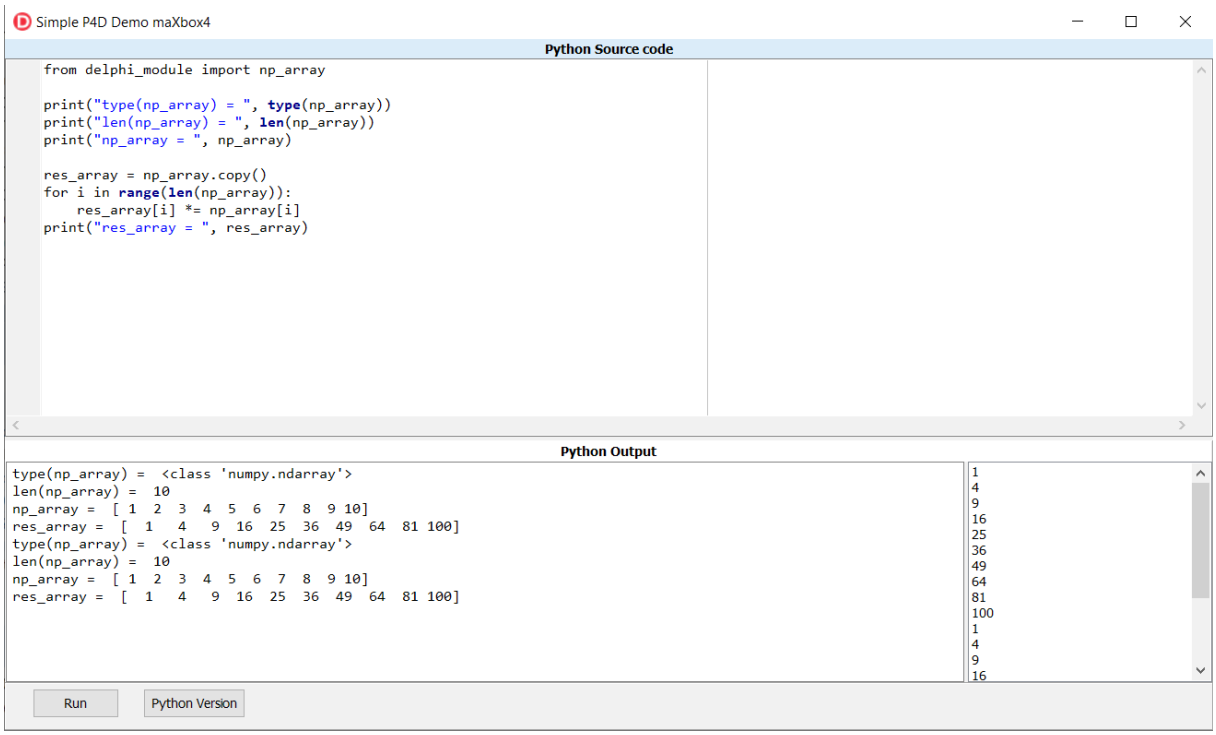
## 1.3 Scatter Plot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

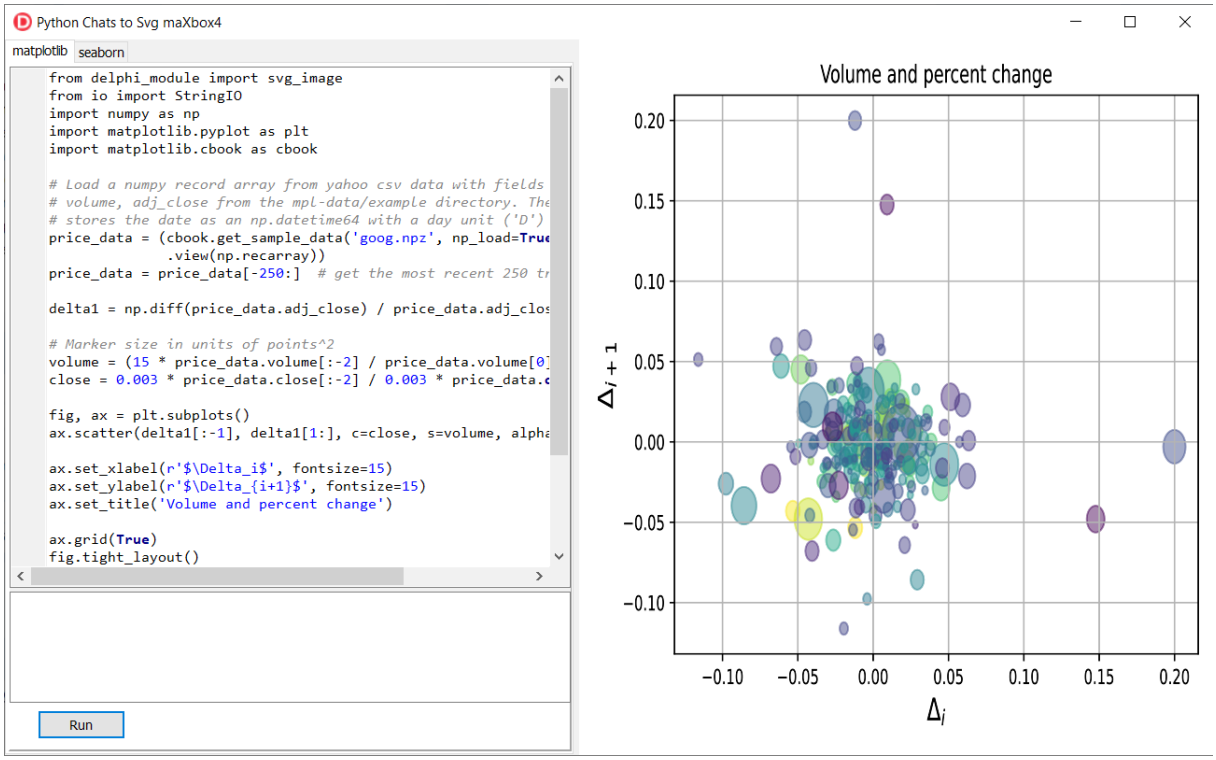
A common modification of the basic scatter plot is the addition of a third variable. Values of the third variable can be encoded by modifying how the points are plotted. For a third variable that indicates categorical values (like geographical region or gender), the most common encoding is through point color.
Seaborn lets you create relational plots using the `relplot()` function. The function technically lets you create more than scatter plots.

Numpy arrays are a good substitute for python lists. They are [better than python lists](). They provide faster speed and take less memory space. Let's begin with its definition for those unaware of numpy arrays. They are multi-dimensional matrices or lists of fixed size with similar elements.

`Pandas` is a popular Python library used to manipulate tabular data. It provides a versatile `dataframe` object that can read data from many popular formats, such as Excel, SQL, CSV and more. The Pandas style API provides you with many different tools that makes working with styling tabular data much easier.



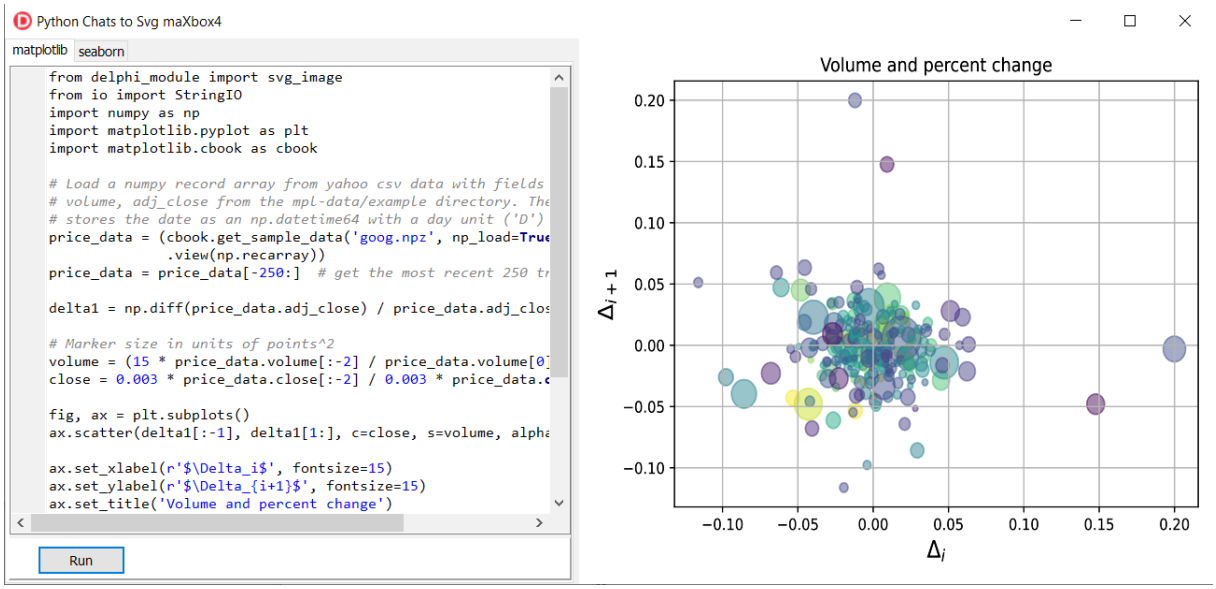

1096_2022-11-12_var_python.png




1096_2022-11-12_svg_python.png

## [python - Scaling a plot (matplotlib) - Stack Overflow]

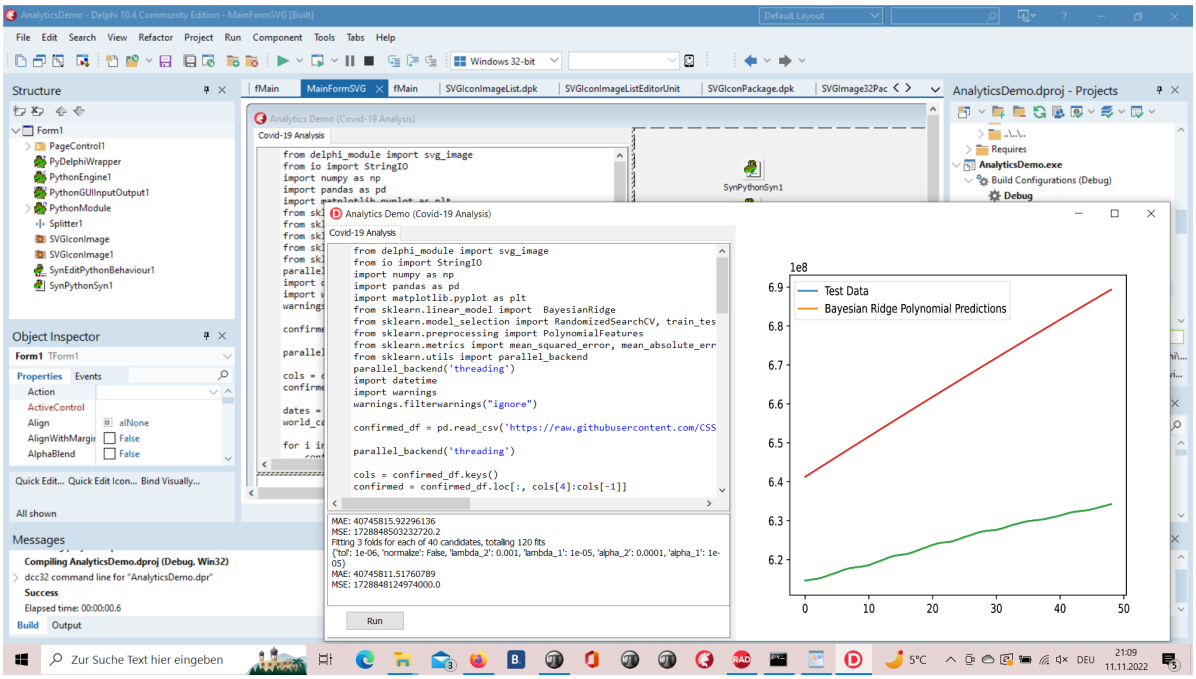
The graph is quite big, so in order to not only see many overlapping dots indicating the nodes, I have to scale the output picture. I used: f,ax = plt.subplots (1,1) ax.set_aspect ('equal') zoom=30 w, h = f.get_size_inches () f.set_size_inches (w * zoom, h * zoom) .




1096_2022-11-12_svg_python2.png

## 1.4 Predict and Dict the Pic

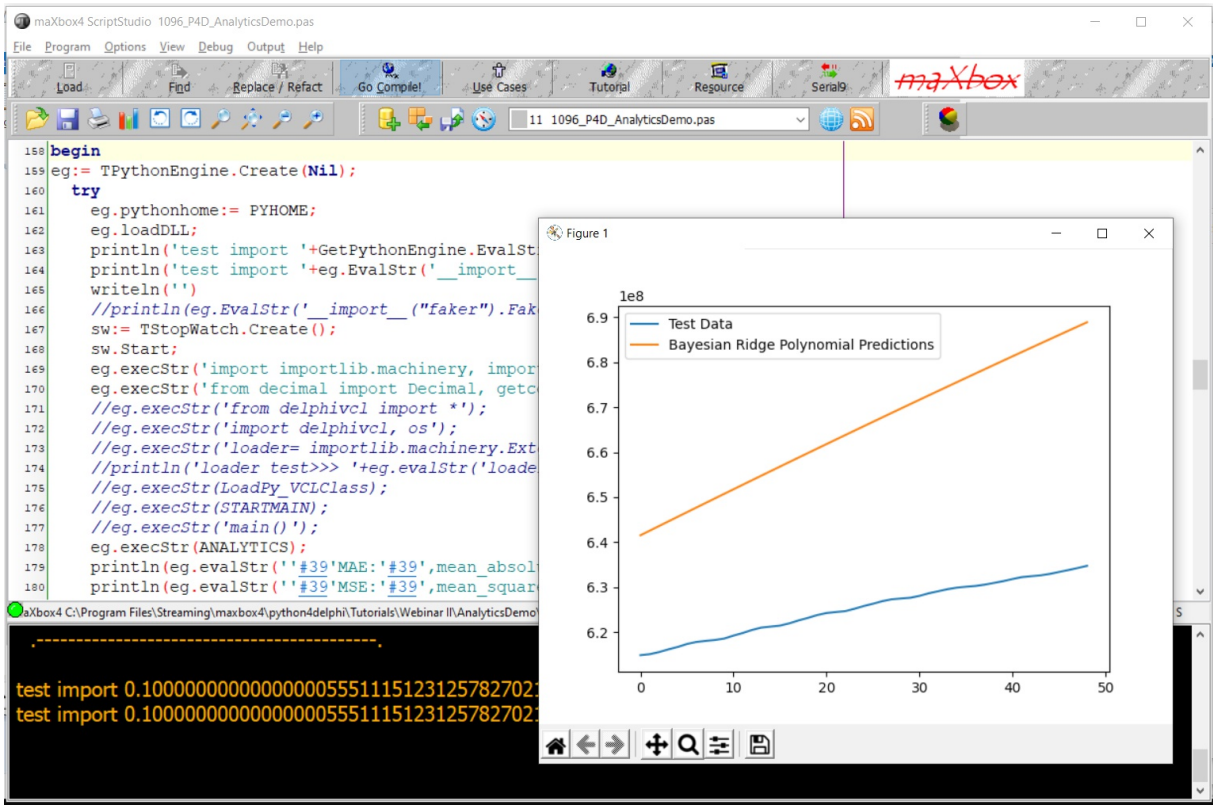
When you hear the word, 'Bayesian', you might think of Naive Bayes. However, Bayesian principles can also be used to perform regression. In this article, we will discuss and implement **Bayesian Ridge Regression**, which is not the same as regular Ridge Regression. To understand more about regular Ridge Regression, you can follow [this] link.




**1096_2022-11-11_analytics.png**

## 1.5 Predict in a Script

Regression is a Machine Learning task to predict continuous values (real numbers), as compared to classification, that is used to predict categorical (discrete) values.




1096_2022-11-11_analytics_script.png

The algorithm uses a hyperparameter to control regularization strength and fully integrates over the hyperparameter in the posterior distribution, applying a hyperprior selected so as to be approximately noninformative.

Scikit-learn's algorithm makes use of conjugate priors and because of that is restricted to use the Gamma prior which requires four hyperparameters chosen arbitrarily to be small values. Additionally, it requires initial values for parameters αα and λλ that are then updated from the data.

```
import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear_model import BayesianRidge
```

In comparison, the algorithm we presented requires no initial parameters; and because the hyperparameter is integrated over, poor performing values contribute little to the posterior probability mass.

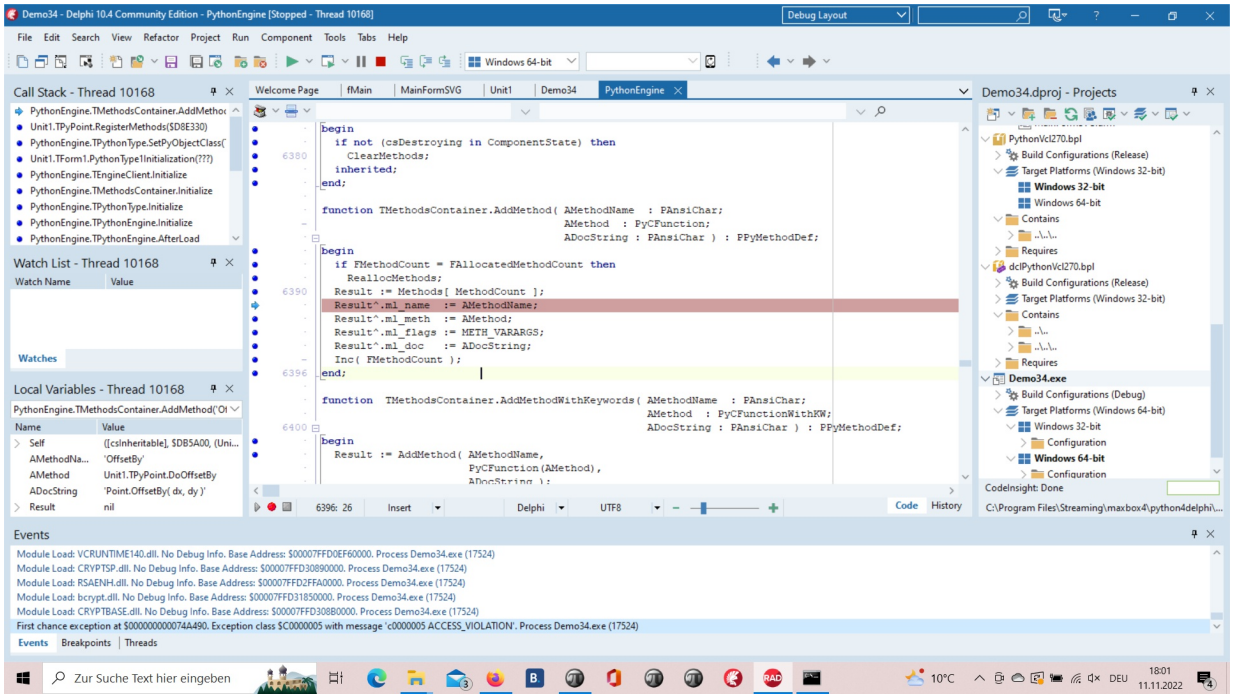
```
# ##############################################################################
# Plot the true and predicted curves for bbai's BayesianRidgeRegression
model
from bbai.glm import BayesianRidgeRegression
reg = BayesianRidgeRegression(fit_intercept=False)
fig, ax = plt.subplots(1, 1, figsize=(4, 4))
```

```
# Note: there are no parameters to tweak
reg.fit(X_train, y_train)
ymean, ystd = reg.predict(X_test, return_std=True)

ax.plot(x_test, func(x_test), color="blue", label="sin($2\\pi x$)")
ax.scatter(x_train, y_train, s=50, alpha=0.5, label="observation")
ax.plot(x_test, ymean, color="red", label="predict mean")
ax.fill_between(
    x_test, ymean - ystd, ymean + ystd, color="pink", alpha=0.5,
label="predict std"
)
ax.set_ylim(-1.3, 1.3)
ax.legend()

plt.tight_layout()
plt.show()
```

# 2 Technical Description

**maxkleiner (Max Kleiner) (github.com)**

A Python program terminates as soon as it encounters an error. In Python, an error can be a syntax error or an exception. In this article, you will see what an exception is and how it differs from a syntax error. After that, you will learn about raising exceptions and making assertions. Then, you'll finish with a demonstration of the try and except block.
https://realpython.com/python-exceptions/



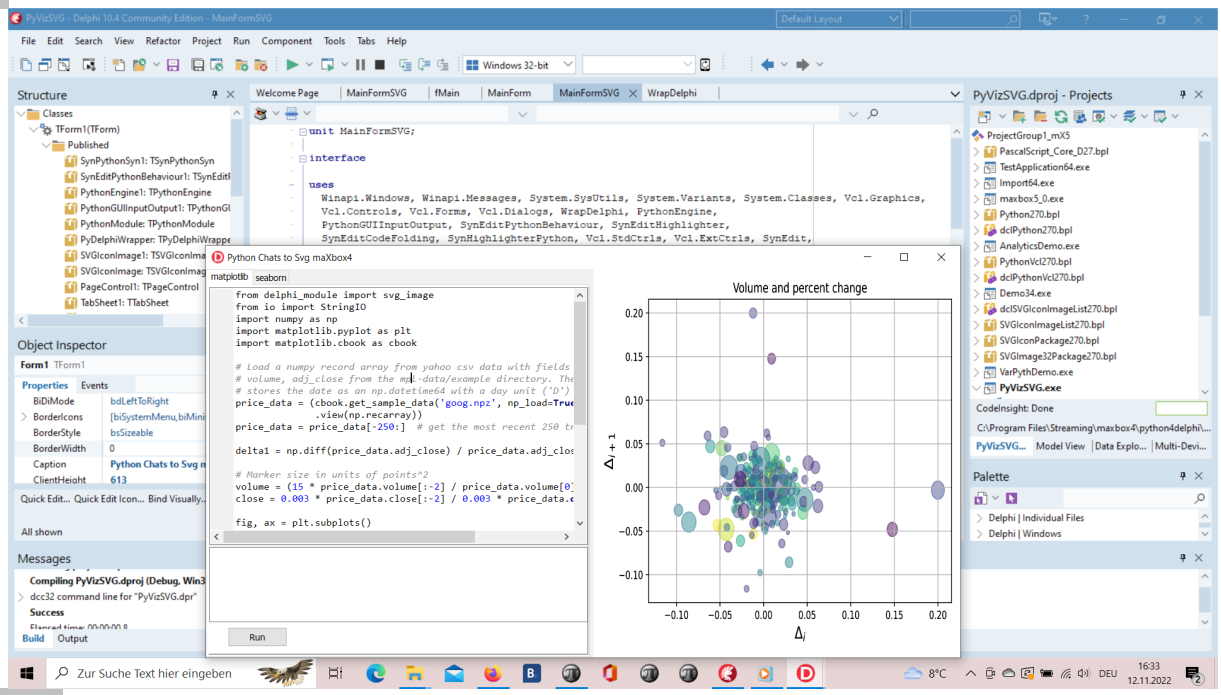

This Exception was prior to change between 32- and 64 bit.

**1096_2022-11-11_exception.png**

# 3 Develop Overview

**We value the quality of the material each student is learning. Our Instructor led classes help provide hands-on learning, one-on-one mentoring with experienced developers and peer-to-peer**

**learning. To allow the most flexibility for students, we have Day and Night programs to choose from. *No prior coding experience required.**




**1096_2022-11-12_develop_overview.png**

```
procedure pyBank_VCL4Python;
var eg: TPythonEngine; sw: TStopWatch;
begin
 eg:= TPythonEngine.Create(Nil);
  try
    eg.pythonhome:= PYHOME;
    eg.loadDLL;
    println('test import
'+GetPythonEngine.EvalStr('__import__("decimal").Decimal(0.1)'));
    writeln('')
    //println(eg.EvalStr('__import__("faker").Faker()'));
    sw:= TStopWatch.Create();
    sw.Start;
    eg.execStr('import importlib.machinery, importlib.util');
    eg.execStr('from decimal import Decimal, getcontext');
    importlib.machinery.ExtensionFileLoader("DelphiVCL",'+VCLHOME+')')

    eg.execStr(ANALYTICSSVG);

//println(eg.evalStr(''#39'MSE:'#39',mean_squared_error(test_bayesian_pred,
y_test_confirmed)'));

    sw.Stop;
    //sw.ElapsedMilliseconds;
    writeln('Stop Analytics Tester1: '+sw.getValueStr)
  except
    eg.raiseError;
    writeln(ExceptionToString(ExceptionType, ExceptionParam));
  finally
    eg.Free;
    sw.Free;
    sw:= Nil;
  end;
end;
```