Machine Learning Plus
(https://www.machinelearningplus.com/)
Let's Data Science

Python Boxplot

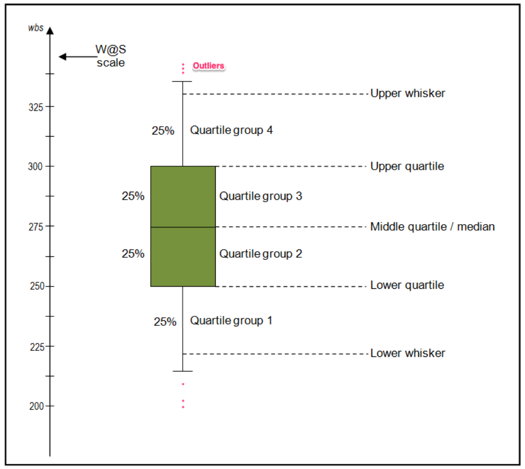
by Venmani A D (https://www.machinelearningplus.com/author/venmani/) |

Boxplot is a chart that is used to visualize how well a dataset is distributed using quartiles. It shows the minimum, maximum, median, first quartile and third quartile in the data set.

What is a boxplot?

Box plot is way to graphically depict groups of numerical data through their quartiles.

From the below image you can see what information we generally get from a box plot.



(https://www.machinelearningplus.com/wp-content/uploads/2020/04/representation.png)

How to interpret the box plot?

The bottom of the (green) box is the 25% percentile and the top is the 75% percentile value of the data.

So, essentially the box represents the middle 50% of all the datapoints which represents the core region when the data is situated. The height of the boxplot is also called the Inter Quartile Range (IQR), which mathematically is the difference between the 75th and 25th percentile values of the data.

Recent Posts

Requests in Python (Guide)
(https://www.machinelearningplus.com/python/req
in-python/)

Matplotlib Pyplot
(https://www.machinelearningplus.com/plots/matp
pyplot/)

Python Boxplot
(https://www.machinelearningplus.com/plots/pythc
boxplot/)

Bar Plot in Python
(https://www.machinelearningplus.com/plots/bar-plot-in-python/)

data.table in R – The Complete Beginners Guide
(https://www.machinelearningplus.com/data-manipulation/datatable-in-r-complete-guide/)

Augmented Dickey Fuller Test (ADF Test) – Must Read Guide
(https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/)

KPSS Test for Stationarity
(https://www.machinelearningplus.com/time-series/kpss-test-for-stationarity/)

101 R data.table Exercises
(https://www.machinelearningplus.com/data-manipulation/101-r-data-table-exercises/)

P-Value – Understanding from Scratch
(https://www.machinelearningplus.com/statistics/p-value/)

101 Python datatable Exercises (pydatatable)
(https://www.machinelearningplus.com/data-manipulation/101-python-datatable-exercises-pydatatable/)

Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python
(https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/)

Mahalonobis Distance – Understanding the math with examples (python)
(https://www.machinelearningplus.com/statistics/m
distance/)

datetime in Python – Simplified Guide with Clear Examples
(https://www.machinelearningplus.com/python/dat
python-examples/)

The thick line in the middle of the box represents the median. Whereas, the upper and lower whisker marks 1.5 times the IQR.

But, why whiskers matter?

Because, the points that lie outside the whiskers, that is, (1.5 x IQR) in both directions are generally considered as outliers.

Lets create an artificial dataset and visualize the data using box plot. For creating an artificial dataset I used random.rand() command from numpy which generates random values between 0-1 to each element in the array.

And I speciified the length of the Dataframe to be 50×3 as argument in the function

```
 import pandas as pd
import numpy as np
df = pd.DataFrame(np.random.rand(50, 3), columns=['A', 'B', 'C'])
df.head()
```

	A	B	C
0	0.160039	0.288336	0.489377
1	0.947344	0.897678	0.356333
2	0.486124	0.928932	0.263846
3	0.022501	0.700835	0.298088
4	0.742429	0.503843	0.296938

The dataset contains 50 randomly selected values between 0-1 in each column.

Looking into this data and finding it's distribution will take an ample amount of time, thats where using a distribution plot like boxplot comes in handy.

Basic boxplot using pandas library

Since, we are dealing with a pandas dataframe, you can create the boxplot using the `pandas` library directly.

`df` is the DataFrame we cretaed before, for plotting boxplot we use the command `DataFrame.plot.box()` .

```
# Boxplot with Pandas
df.plot.box(title='Boxplot with pandas');
```

Feedback

Principal Component Analysis (PCA) – Better Explained
(https://www.machinelearningplus.com/machine-learning/principal-components-analysis-pca-better-explained/)

Python Logging – Simplest Guide with Full Code and Examples
(https://www.machinelearningplus.com/python/pyt
logging-guide/)

Matplotlib Histogram – How to Visualize Distributions in Python
(https://www.machinelearningplus.com/plots/matp
histogram-python-examples/)

ARIMA Model – Complete Guide to Time Series Forecasting in Python
(https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

Time Series Analysis in Python – A Comprehensive Guide with Examples
(https://www.machinelearningplus.com/time-series/time-series-analysis-python/)

Matplotlib Tutorial – A Complete Guide to Python Plot w/ Examples
(https://www.machinelearningplus.com/plots/matp
tutorial-complete-guide-python-plot-examples/)

Topic modeling visualization – How to present the results of LDA models?
(https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/)

Top Posts & Pages

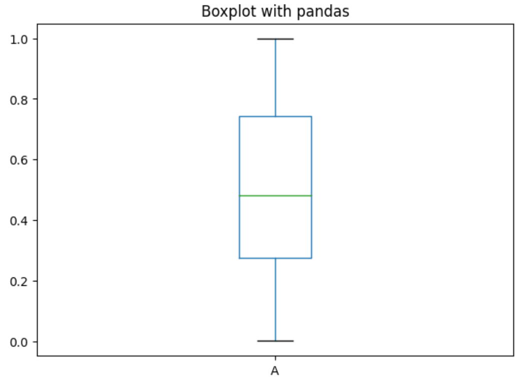
ARIMA Model - Complete Guide to Time Series Forecasting in Python
(https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)

Time Series Analysis in Python - A Comprehensive Guide with Examples
(https://www.machinelearningplus.com/time-series/time-series-analysis-python/)

Top 50 matplotlib Visualizations - The Master Plots (with full python code)
(https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/)

Cosine Similarity - Understanding the math and how it works (with python codes)
(https://www.machinelearningplus.com/nlp/cosine-similarity/)

Parallel Processing in Python - A Practical Guide with Examples
(https://www.machinelearningplus.com/python/par
processing-python/)



[https://www.machinelearningplus.com/wp-content/uploads/2020/04/simple-boxplot.png]

From this you can see the median value for each distribution and also the different quartile groups.

Plotting a boxplot using matplotlib

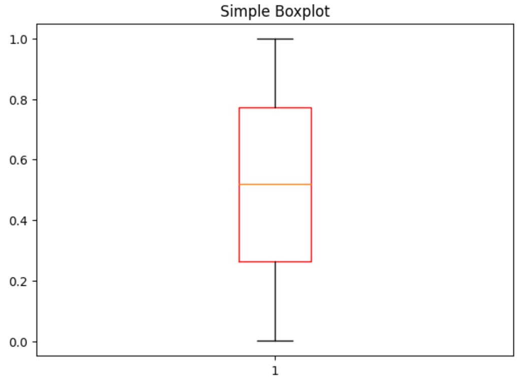
For using matplolib, first you need to import the matplotlib library. If you want you can alter the default parameters by using the function plt.rcParams.update() function.

Then, use plt.boxplot(data) for plotting the data.

```
# Load package
import pandas as pd, numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
plt.rcParams.update({'figure.figsize':(7,5), 'figure.dpi':100})

# Creating dataset
df = pd.DataFrame(np.random.rand(500, 1), columns=['A'])

# plot
plt.boxplot(df['A'], boxprops=dict(color='red'))
plt.title('Simple Boxplot');
```

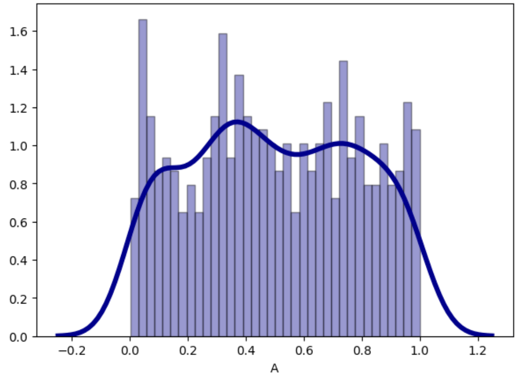


[https://www.machinelearningplus.com/wp-content/uploads/2020/04/matplotlib-boxplot.png]

Alternately, you can visualize the distribution with a combination of histogram and density plot. This is straightforward to create with seaborn .

```
# Histogram and density
import seaborn as sns

sns.distplot(df['A'], hist=True, kde=True,
             bins=int(180/5), color = 'darkblue',
             hist_kws={'edgecolor':'black'},
             kde_kws={'linewidth': 4});
```



[https://www.machinelearningplus.com/wp-content/uploads/2020/04/normal-density-plot.png]

Notched Boxplot in matplotlib

The notched boxplot allows you to evaluate confidence intervals (by default 95% confidence interval) for the medians of each boxplot.

To create the notch, set notch=True in the plt.boxplot function.

Topic Modeling with Gensim (Python) [https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/]

Feedback

101 Pandas Exercises for Data Analysis [https://www.machinelearningplus.com/python/101-pandas-exercises-python/]

Complete Introduction to Linear Regression in R [https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/]

Matplotlib Histogram - How to Visualize Distributions in Python [https://www.machinelearningplus.com/plots/matp histogram-python-examples/]

101 NumPy Exercises for Data Analysis (Python) [https://www.machinelearningplus.com/python/101-numpy-exercises-python/]

Tags

Bigrams [https://www.machinelearningplus.com/tag/bigrams/] Classification [https://www.machinelearningplus.com/tag/c Corpus [https://www.machinelearningplus.com/tag/corpus/] Cosine Similarity [https://www.machinelearningplus.com/tag/cosine-similarity/] data.table [https://www.machinelearningplus.com/tag/data-table/] Data Manipulation [https://www.machinelearningplus.com/ta manipulation/] Debugging [https://www.machinelearningplus.com/tag/debugging/] Doc2Vec [https://www.machinelearningplus.com/tag/doc2vec/] Evaluation Metrics [https://www.machinelearningplus.com/tag/evaluation-metrics/] FastText [https://www.machinelearningplus.com/tag/fasttext/] Feature Selection [https://www.machinelearningplus.com/tag/feature-selection/] Gensim [https://www.machinelearningplus.com/tag/gensi LDA [https://www.machinelearningplus.com/tag/lda/] Lemmatization [https://www.machinelearningplus.com/tag/lemmatization/] Linear Regression [https://www.machinelearningplus.com/tag/linear-regression/] Logistic [https://www.machinelearningplus.com/tag/logistic/] LSI [https://www.machinelearningplus.com/tag/lsi/] Matplotlib [https://www.machinelearningplus.com/ Multiprocessing [https://www.machinelearningplus.com/tag/multiprocessing/] NLP [https://www.machinelearningplus.com/tag/n NLTK [https://www.machinelearningplus.com/tag/nltk/] Numpy [https://www.machinelearningplus.com/tag/nump P-Value [https://www.machinelearningplus.com/tag/p-value/] Pandas [https://www.machinelearningplus.com/tag/pandas/] Parallel Processing

Feedback

[https://www.machinelearningplus.com/tag/parallel-processing/] Phraser [https://www.machinelearningplus.com/tag/phraser/] Practice Exercise [https://www.machinelearningplus.com/tag/practice-exercise/] Python [https://www.machinelearn R [https://www.machinelearningplus.cc Regex [https://www.machinelearningplus.com/tag/regex/] Regression [https://www.machinelearningplus.com/tag/regression/] Residual Analysis [https://www.machinelearningplus.com/tag/residual-analysis/] Scikit Learn [https://www.machinelearningplus.com/tag/scikit-learn/] Significance Tests [https://www.machinelearningplus.com/tag/signifi tests/] Soft Cosine Similarity [https://www.machinelearningplus.com/tag/soft-cosine-similarity/] spaCy [https://www.machinelearningplus.com/tag/spacy/] Stationarity [https://www.machinelearningplus.com/tag/statio Summarization [https://www.machinelearningplus.com/tag/summarization/] TaggedDocument [https://www.machinelearningplus.com/tag/taggeddocument/] TextBlob [https://www.machinelearningplus.com/tag/textblob/] TFIDF [https://www.machinelearningplus.com/tag/tfidf/] Time Series [https://www.machinelearningplus.com/ta series/] Topic Modeling [https://www.machinelearningplus.com/tag/topic-modeling/] Visualization [https://www.machinelearningplus.com/tag/visual Word2Vec [https://www.machinelearningplus.com/tag/word2vec/]