

**MULTIVARIATE ANALYSIS AND MACHINE LEARNING TECHNIQUES- FEATURE ANALYSIS IN DATA SCIENCE USING PYTHON**

**DATA SETS AND SOFTWARE PACKAGES**



Version 2.1

Table of Contents

[1 DESCRIPTION OF DATA SETS 2](#_Toc125976865)

[1.1 Bank Marketing 2](#_Toc125976866)

[1.2 Breast Cancer - Gene Prognostic Signature 2](#_Toc125976867)

[1.3 ChickWeight R Documentation 2](#_Toc125976868)

[1.4 CIFAR-10 3](#_Toc125976869)

[1.5 CIFAR-100 3](#_Toc125976870)

[1.6 diamonds R Documentation 3](#_Toc125976871)

[1.7 German Breast Cancer Study Group 2 (GBSC2) 4](#_Toc125976872)

[1.8 iris R Documentation 4](#_Toc125976873)

[1.9 IT Projects 4](#_Toc125976874)

[1.10 MNIST handwritten character recognition 4](#_Toc125976875)

[1.11 Penguins 5](#_Toc125976876)

[1.12 PowerConsumption\_2013\_19 5](#_Toc125976877)

[1.13 Tips 5](#_Toc125976878)

[1.14 West Virginia Student Fraternity 5](#_Toc125976879)

[2 SOFTWARE PACKAGE INSTALLATION 6](#_Toc125976880)

[2.1 Package Managers – conda, and pip 6](#_Toc125976881)

[2.2 Install Anaconda 7](#_Toc125976882)

[2.3 Modify the ‘path’ variable 8](#_Toc125976883)

[2.4 Install Python 8](#_Toc125976884)

[2.5 Launch Spyder IDE for program development 8](#_Toc125976885)

[2.6 Install Other Packages 9](#_Toc125976886)

[2.7 Install TensorFlow and Keras 9](#_Toc125976887)

[2.8 Notes on Google Colab, Spyder, and other IDEs 10](#_Toc125976888)

[2.9 Check / Resolve conflicts in package installation 10](#_Toc125976889)

[2.10 Installation from the Command Prompt - for experienced programmers 11](#_Toc125976890)

# DESCRIPTION OF DATA SETS

Note that some of the following datasets are available in this repository.

## Bank Marketing

Refer – https://archive.ics.uci.edu/ml/datasets/bank+marketing

The dataset consists of data gathered from a bank marketing survey. It has 21 variables and 45211 rows. Feature variables considered in the exercises are 'age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous'. The target (independent variable) is ‘Has the client in the training dataset, subscribed to a term deposit?’ with binary answers ('yes’, ‘no').

## Breast Cancer - Gene Prognostic Signature

Refer:- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7390

A Survival analysis study was conducted on 198 breast cancer patients to predict distant metastases (dm). Distant metastasis (dm) refers to cancer that has spread from the original (primary) tumor to distant organs/lymph nodes. The study included 76-gene prognostic signatures in lymph node-negative (N-) breast cancer patients.

## ChickWeight R Documentation

The ChickWeight data frame has 578 rows and 4 columns from an experiment on the effect of diet on early growth of chicks.The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets.

weight: a numeric vector giving the body weight of the chick (gm).

Time: a numeric vector giving the number of days since birth when the measurement was made.

Diet: a factor with levels 1, ..., 4 indicating which experimental diet the chick received.

Chick: {not used in the tutorials} an ordered factor with levels 18 < ... < 48 giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.

**Source**

Crowder, M. and Hand, D. (1990), Analysis of Repeated Measures, Chapman and Hall (example 5.3)

Hand, D. and Crowder, M. (1996), Practical Longitudinal Data Analysis, Chapman and Hall (table A.2)

Pinheiro, J. C. and Bates, D. M. (2000) Mixed-effects Models in S and S-PLUS, Springer.

## CIFAR-10

Refer – https://www.cs.toronto.edu/~kriz/cifar.html

The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class.

## CIFAR-100

Refer – https://www.cs.toronto.edu/~kriz/cifar.html

CIFAR-100 dataset has 100 classes containing 600 images each. The 100 classes are grouped into 20 superclasses.

## diamonds R Documentation

A dataset containing the prices and other attributes of almost 54,000 diamonds.

**Refer**: https://www.gemsociety.org/article/diamond-color/

The variables are: -

diamonds: A data frame with 53940 rows and 10 variables:

price: price in US dollars (\$326–\$18,823)

carat: weight of the diamond (0.2–5.01)

cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color: diamond colour, from D (best) to J (worst)

clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x: length in mm (0–10.74)

y: width in mm (0–58.9)

z: depth in mm (0–31.8)

depth: total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43–79)

table: width of top of diamond relative to widest point (43–95)

## German Breast Cancer Study Group 2 (GBSC2)

**Refer**: https://ascopubs.org/doi/abs/10.1200/jco.1994.12.10.2086

A dataset with having 686 samples and 8 features. The endpoint is recurrence free survival, which occurred for 299 patients (43.6%). This dataset is the part of a study by the German Breast Cancer Study Group (GBSG) in 1984. This was a randomized clinical trial to compare the effectiveness of different drugs dosage, on recurrence-free and overall survival

## iris R Documentation

The data frame iris has 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

It has measurements in centimeters of sepal/ petal length and width, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

**Source**

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179–188.

The data were collected by Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2–5.

**Refer**

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole. (has iris3 as iris.)

## IT Projects

**Refer:**

https://www.springerprofessional.de/en/variation-of-risk-profile-across-software-life-cycle-in-is-outso/17079520

A study was conducted to explore the variables that influence software project performance, a float. Eleven variables were hypothesized to influence project performance – they are measured on a scale of 0..6. 100 responses were collected from IT consultants working on various projects. Project Type is a category Variable with three labels - 0,1,2

## MNIST handwritten character recognition

**Refer**: http://yann.lecun.com/exdb/mnist/

The MNIST database of handwritten digits houses 60,000 training samples and 10,000 test samples of handwritten digit images in black and white {0,1}. The digits have been size-normalized and centered in a fixed-size image of 28x28. Develop a convolutional neural network to classify the images. Describe the network structure and parameters

## Penguins

**Refer:** https://github.com/allisonhorst/penguins

The dataset contain data for 344 penguins. There are 3 different species of penguins in this dataset, collected from 3 islands in the Palmer Archipelago, Antarctica.

## PowerConsumption\_2013\_19

**Refer:** https://github.com/srikrishnan-punalur/MVA-ML

The dataset consists of data gathered from the electricity board, for a state in India, reported from 2013 to 2019. The features include daily power demand, shortage, consumption, load, OD/UD, and date. These are typical time-series / sequences and can be used for RNN/ LSTM exercises.

## Tips

**Refer:**

https://github.com/mwaskom/seaborn-data/blob/master/tips.csv

https://www.kaggle.com/ranjeetjain3/seaborn-tips-dataset

Food servers’ tips in restaurants may be influenced by many factors. The following dataset shows customers served during two and a half months.

## West Virginia Student Fraternity

**Refer:** https://github.com/srikrishnan-punalur/MVA-ML

This dataset consists of information about the interaction of students in a student fraternity in West Virginia - the number of times they interacted over a week. This dataset is used for social network analysis in Chapter 10.

# SOFTWARE PACKAGE INSTALLATION

Software libraries get revised frequently. Therefore, the commands, and software versions mentioned below may undergo changes. Refer to the home page of the respective software projects for current usage. The topics covered in this section are: -

1. Package Managers – conda, and pip

2. Software Package Installation

2.1 Install Anaconda

2.2 Modify the ‘path’ variable

2.3 Install Python

2.4 Launch Spyder IDE for program development

2.5 Install Other Packages

2.6 Install TensorFlow and Keras

3. Notes on Google Colab, Spyder, and other IDEs

4. Check / Resolve conflicts in package installation

5. Installation from the Command Prompt - for experienced programmers

## Package Managers – conda, and pip

Conda and pip are both used for installing Python packages. We can install anaconda and use the conda command from the anaconda prompt to install other packages of the Python ecosystem. Alternatively, the pip command from the command prompt or the !pip from a Python IDE can be used to install packages. The earlier method (conda) is recommended for beginners, though the latter method may be used by professionals for installing specific packages and setting up work environments. The objective of this section is to develop a primary understanding of software installation for building a working Python development environment for machine learning.

As mentioned above, we can use the !pip for package installation in Google Colab, and other IDEs. The command the !pip is also used to check and resolve path dependencies of the installed packages. This ensures that the packages are not broken and are accessible for use.

The Python Package Index (PyPI) is a repository for Python Software packages. PyPI. Python software package developers use PyPI to distribute their software. PyPI has a massive 348,665 projects. **Pip** is the Python Packaging Authority’s recommended package manager for installing python packages from the Python Package Index. Pip installs Python software packaged as wheels or source distributions. The latter may require that the system have compatible compilers, and possibly libraries, installed before invoking pip to succeed.

Conda is a cross-platform package and environment manager that installs and manages software packages from the Anaconda repository and the Anaconda Cloud. The repository has about 1500 software packages. They have Python packages and other software such as C or C++ libraries, and R packages. Conda packages are binaries. Therefore, there is no need for compilation on the host system.

Python must be installed, before using pip manager. We may download Python from python.org and install it. However, Conda can install Python, and Python packages directly using conda commands. Pip depends on other tools like virtualenv or venv to create isolated environments. However, Conda can create isolated environments that consist of different versions of Python/packages.

While installing packages, pip installs dependencies in a recursive, serial loop. This may lead to environments that are broken if the compatibility of packages installed earlier conflicts with the packages installed later. However, conda checks the package metadata and resolves dependencies, thereby creating a robust work environment. The steps involved in software installation are shown below.

## Install Anaconda

It is recommended that we download the latest version of Anaconda and install it before installing Python or Python packages. While installation progresses, be attentive; and answer the questions (like proceed Yes/No). Note the following points:-

* If we accept the default option to install Anaconda, it is installed in your user home directory:

C:\Users\<your-username>\Anaconda3\

* It is recommended that we do not install anaconda in paths that contain spaces such as C:\Program Files. Instead, create your directory in an appropriate folder with an appropriate name. For example,

C:\anaconda

* For doing the above (a) you must select the ‘multiuser’ option and (b) you must have admin authority.
* While installing Anaconda, it is recommended that you may not add the anaconda path (C:\anaconda) to the Windows PATH.

## Modify the ‘path’ variable

The OS uses certain variables, one of which is ‘path’. This variable contains a list of directories where the system will search for executable code. We need to add the path where anaconda scripts are present.

* Modify the windows path by editing 'environment variables' (from the start menu)
* Click on ‘path’ variable
* Click ‘edit’
* Add a 'new' path - the place where anaconda scripts are installed – e.g., it may look like the one below

C:\Users\<Your Usernae>\anaconda3\Scripts\

* Click ok
* Now the 'path' variable is updated. The system knows where to look for anaconda scripts!

## Install Python

After installing anaconda, you may install Python, using Conda prompt. For this, on Windows, open the Start menu and open Anaconda Command Prompt. Execute the following command

conda create -n py39 python=3.9

conda activate py39

(assuming that python 3.9 is the latest stable version of python)

If an older version of Python is already installed, you may upgrade it from the ‘anaconda prompt’ using the above command.

## Launch Spyder IDE for program development

In windows start menu, click on 'anaconda navigator' app. When it opens up, launch 'Spyder' inside 'anaconda navigator'. Now Spyder IDE is available for programming

**Spyder – font size, theme**

Open Spyder.

Select 'Tools' menu.

Then select submenu 'preferences'.

Select 'appearance'

Here you may change font size (fonts -size)

You may also change background and text color layout (icon theme --> 'spyder')

Now click 'OK'

**Spyder – scale**

You may again select 'Tools' menu. Then select submenu 'preferences'.

You may select 'application' to change the scale of display

## Install Other Packages

We can use the anaconda package manager to install any package. See Table 1-1.

Table ‑ Using the ‘conda’ command for Package Installation

|  |  |
| --- | --- |
| Run ‘conda’ Package Manager Command | Package Functionality  (See Table1-2 for references |
| conda install -c conda-forge python-graphviz  conda install -c conda-forge pingouin  conda install -c desilinguist factor\_analyzer  conda install -c sebp scikit-survival  conda install -c conda-forge eli5  {Use *pip install apyori* command; See Table 1-2} | To Draw Decision Tree  Partial Correlation coef.  Factor Analysis  Survival Analysis  Survival -Feature Ranking  Market Basket Analysis |

* if a package is already installed, it may be upgraded using the same command used above. For example,

conda install pingouin

## Install TensorFlow and Keras

On Windows, open the Start menu, then open an Anaconda Command Prompt.

* Choose a name for your TensorFlow environment, such as ‘TensorFlow’. Beginners are recommended to install the current release of CPU-only TensorFlow.

conda create -n tf tensorflow

conda activate tf

* After installing anaconda, python, and tensor flow, install Keras

conda install -c conda-forge keras

(or Run conda install -c conda-forge/label/broken keras)

## Notes on Google Colab, Spyder, and other IDEs

We saw how to use conda or pip for package installation. There is yet another method. We can use the !pip command from within an IDE (e.g., google colab or sypder) for package installation. See Table 1-2. So, how do you install python packages, in Google Colab? You must use the above method (!pip command) from within the Colab.

Table ‑ !pip - Google Colab and other IDEs, Package Installation / Path Resolution

|  |  |
| --- | --- |
| Installation of Python Packages in Google Colab or other IDEs (the same can be used to resolve code break issues, due to incorrect installation) | |
| !pip3 install --upgrade pip | <https://pypi.org/project/pip/> |
| !pip install graphviz | <https://graphviz.org/> |
| !pip install pingouin | https://pingouin-stats.org/ |
| !pip install factor-analyzer | <https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html> |
| !pip install apyori | <https://pypi.org/project/apyori/> |
| !pip scikit-survival | <https://scikit-survival.readthedocs.io/en/stable/install.html> |
| !pip eli5 | (Same as above) |
| !pip3 install --upgrade tf | <https://www.tensorflow.org/install> |

## Check / Resolve conflicts in package installation

When we try to access an installed package from within an IDE, we may get the error - ‘ModuleNotFoundError’. One of the reasons may that the system is not able to resolve the access path. !pip command can be used to check and resolve path dependencies of the installed packages (See Table 1-2). This ensures that the packages are not broken and are accessible for usage. You may do the following inspections to ensure smooth library access.

* To check whether a software package properly installed
* Let’s check ‘FactorAnalyzer’ package
* In the Syder IDE, enter the following command and ‘run’ it

from factor\_analyzer import FactorAnalyzer

>> ModuleNotFoundError: No module named 'factor\_analyzer'

* Now let’s remedy the error. Enter the following commands and ‘run’ it

!pip install factor\_analyzer

>> Successfully installed ...

from factor\_analyzer import FactorAnalyzer

* Check TensorFlow installation

!python -m pip show tf

>> WARNING: Package(s) not found: tf

* TensorFlow – Resolve the error reported above

!pip3 install --upgrade tf

!python -m pip show tf

>> Name: tf

>> Version: 1.0.0

>> Location: c:\users\sri-s\anaconda3\lib\site-packages

* Check Keras installation

!python -c "import keras; print(keras.\_\_version\_\_)"

2.9.0

* Check Access to any Package, from within an IDE

Import pandas as pd

from sksurv.datasets import load\_gbsg2

## Installation from the Command Prompt - for experienced programmers

It is recommended that we use the conda package manager for installing the Python work environment. However, we may need to use the pip package manager from the command prompt for some packages. See Table 1-3. The only difference with Table 1-2, is the absence of the exclamation mark (pip, instead of !pip)

Table ‑ System Command Prompt pip Command and package references

|  |  |
| --- | --- |
| Installing Python Packages from system command prompt | Download the package to an appropriate directory. From there issue pip command |
| pip3 install --upgrade pip | Package Manager pip |
| pip install graphviz | <https://graphviz.org/> |
| pip install pingouin | https://pingouin-stats.org/ |
| pip install factor-analyzer | [https://factor-analyzer.readthedocs.io/en/latest/factor\_analyzer.html](https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html%20) |
| pip install apyori | https://pypi.org/project/apyori/ |
| pip scikit-survival  pip eli5 | https://scikit-survival.readthedocs.io/en/stable/install.html |

We may download a package from the web, install it, and add the path to the ‘PATH’ variable. For example, download the Graphviz installer to an appropriate folder, e.g., C:\Analytics\graphviz. After downloading the installer, we can use ‘pip’ from the Microsoft Windows command prompt 'cmd'.

* Download-> graphviz-2.50.0 (64-bit) EXE installer [sha256]
* CD C:\Analytics\graphviz
* pip install graphviz
* Check the python version in the IDE

import os

os.environ["PATH"] #check the current access paths

* If graphvz is not present in the path, add it to the path, as follows:-

os.environ["PATH"] += os.pathsep + 'C:\\Analytics\\graphviz\\bin'

* In the same way, as described above, other packages can be installed
* if a package is already installed, it may be upgraded. For example,

pip install --upgrade pingouin

pip install pingouin