

- How to use dataframes in Python?
- Import data set as dataframe
- Inspect data frame and access data
- Produce an overview of data features
- Create data plots using matplotlib

Video Tutorials

Pandas Dataframes - Import and Data Handling

Pandas Dataframes - Visualisation

Introduction

This is our first lesson on learning and understanding dataframes in Python.

The diabetes data set is one of the challenging task.

Challenge: The diabetes data set

Here is a screenshot of the so-called diabetes data set. It is taken from this webpage and it is one of the example data sets used to illustrate machine learning functionality in scikit-learn (Part III and Part IV of the course).

< > ↺ ☐ | 🔒 www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt

| AGE | SEX | BMI | BP | S1 | S2 | S3 | S4 | S5 | S6 | Y |
|-----|-----|------|--------|-----|-------|----|------|--------|-----|-----|
| 59 | 2 | 32.1 | 101 | 157 | 93.2 | 38 | 4 | 4.8598 | 87 | 151 |
| 48 | 1 | 21.6 | 87 | 183 | 103.2 | 70 | 3 | 3.8918 | 69 | 75 |
| 72 | 2 | 30.5 | 93 | 156 | 93.6 | 41 | 4 | 4.6728 | 85 | 141 |
| 24 | 1 | 25.3 | 84 | 198 | 131.4 | 40 | 5 | 4.8903 | 89 | 206 |
| 50 | 1 | 23 | 101 | 192 | 125.4 | 52 | 4 | 4.2905 | 80 | 135 |
| 23 | 1 | 22.6 | 89 | 139 | 64.8 | 61 | 2 | 4.1897 | 68 | 97 |
| 36 | 2 | 22 | 90 | 160 | 99.6 | 50 | 3 | 3.9512 | 82 | 138 |
| 66 | 2 | 26.2 | 114 | 255 | 185 | 56 | 4.55 | 4.2485 | 92 | 63 |
| 60 | 2 | 32.1 | 83 | 179 | 119.4 | 42 | 4 | 4.4773 | 94 | 110 |
| 29 | 1 | 30 | 85 | 180 | 93.4 | 43 | 4 | 5.3845 | 88 | 310 |
| 22 | 1 | 18.6 | 97 | 114 | 57.6 | 46 | 2 | 3.9512 | 83 | 101 |
| 56 | 2 | 28 | 85 | 184 | 144.8 | 32 | 6 | 3.5835 | 77 | 69 |
| 53 | 1 | 23.7 | 92 | 186 | 109.2 | 62 | 3 | 4.3041 | 81 | 179 |
| 50 | 2 | 26.2 | 97 | 186 | 105.4 | 49 | 4 | 5.0626 | 88 | 185 |
| 61 | 1 | 24 | 91 | 202 | 115.4 | 72 | 3 | 4.2905 | 73 | 118 |
| 34 | 2 | 24.7 | 118 | 254 | 184.2 | 39 | 7 | 5.037 | 81 | 171 |
| 47 | 1 | 30.3 | 109 | 207 | 100.2 | 70 | 3 | 5.2149 | 98 | 166 |
| 68 | 2 | 27.5 | 111 | 214 | 147 | 39 | 5 | 4.9416 | 91 | 144 |
| 38 | 1 | 25.4 | 84 | 162 | 103 | 42 | 4 | 4.4427 | 87 | 97 |
| 41 | 1 | 24.7 | 83 | 187 | 108.2 | 60 | 3 | 4.5433 | 78 | 168 |
| 35 | 1 | 21.1 | 82 | 156 | 87.8 | 50 | 3 | 4.5109 | 95 | 68 |
| 25 | 2 | 24.3 | 95 | 162 | 98.6 | 54 | 3 | 3.8501 | 87 | 49 |
| 25 | 1 | 26 | 92 | 187 | 120.4 | 56 | 3 | 3.9703 | 88 | 68 |
| 61 | 2 | 32 | 103.67 | 210 | 85.2 | 35 | 6 | 6.107 | 124 | 245 |
| 31 | 1 | 29.7 | 88 | 167 | 103.4 | 48 | 4 | 4.3567 | 78 | 184 |
| 30 | 2 | 25.2 | 83 | 178 | 118.4 | 34 | 5 | 4.852 | 83 | 202 |
| 19 | 1 | 19.2 | 87 | 124 | 54 | 57 | 2 | 4.1744 | 90 | 137 |

This figure captures only the top part of the data. On the webpage you need to scroll down considerably to view the whole content. Thus, to get an **overview** of the dataset is the first main task in Data Science.

The lesson

- introduces code to read and inspect the data
- works with a specific data frame and extracts some techniques to get an overview
- discusses the concept ‘distribution’ as a way of summarising data in a single figure

To get to know a dataset you need to

- access the data
- check the content
- produce a summary of basic properties

In this lesson we will only look at univariate features where each data column is studied independently of the others. Further properties and bivariate features will be the topic of the next lesson.

Work Through Example

Reading data into a Pandas DataFrame

The small practice data file for this section is called 'everleys_data.csv' and can be downloaded using the link given above in "Materials for this Lesson". To start, please create a subfolder called 'data' in the current directory and put the data file in it. It can now be accessed using the relative path `data/everleys_data.csv` or `data\everleys_data.csv`, respectively.

The file `everleys_data.csv` contains blood concentrations of calcium and sodium ions from 17 patients with Everley's syndrome. The data are taken from a BMJ statistics tutorial. The data are stored as comma-separated values (csv), two values for each patient.

To get to know a dataset, we will use the Pandas package and the Matplotlib plotting. The Pandas package for data science is included in the Anaconda distribution of Python. Check this link for installation instructions to get started.

If you are not using the Anaconda distribution, please refer to these guidelines.

To use the functions contained in Pandas they need to be imported. Our dataset is in 'csv' format, and we therefore need to read it from a csv file. For this, we import the function `read_csv`. This will create a *Pandas dataframe*.

```
1 import sys
2 print(sys.version)
```

```
1 3.10.4 (main, Mar 24 2022, 14:07:25) [GCC 9.4.0]
```

```
1 from pandas import read_csv
```

Executing this code does not lead to any output on the screen. However, the function is now ready to be used. To use it, we type its name and provide the required arguments. The following code should import the Everley's data into your JupyterLab notebook (or other Python environment):

```
1 # for Mac OSX and Linux
2 # (please go to the next cell if using Windows)
3
4 df = read_csv("data/everleys_data.csv")
```

```
1 # please uncomment for Windows
2 # (please go to previous cell if using Mac OSX or Linux)
3
4 # df = read_csv("data\everleys_data.csv")
```

This code uses the `read_csv` function from Pandas to read data from a data file, in this case a file with extension 'csv'. Note that the location of the data file is specified within quotes by the relative

path to the subfolder 'data' followed by the file name. Use the JupyterLab file browser to check that subfolder exists and has the file in it.



The screenshot shows the JupyterLab interface. On the left, the file browser displays the directory structure: / L21_DataFrames / data /. Three CSV files are listed: cervical_cancer.csv, diabetes_data.csv, and everleys_data.csv (which is selected). On the right, a preview window for 'everleys_data.csv' is open, showing a table with 18 rows and 3 columns: an index column, a 'calcium' column, and a 'sodium' column. The delimiter is set to a comma (,).

| | calcium | sodium |
|----|---------------|---------------|
| 1 | 3.4555817E+00 | 1.1269098E+02 |
| 2 | 3.6690263E+00 | 1.2566333E+02 |
| 3 | 2.7899104E+00 | 1.0582181E+02 |
| 4 | 2.9399E+00 | 9.8172772E+01 |
| 5 | 5.42606E+00 | 9.7931489E+01 |
| 6 | 7.1581063E-01 | 1.2085833E+02 |
| 7 | 5.6523902E+00 | 1.128715E+02 |
| 8 | 3.5713201E+00 | 1.1264736E+02 |
| 9 | 4.3000669E+00 | 1.3203172E+02 |
| 10 | 1.3694191E+00 | 1.1849901E+02 |
| 11 | 2.550962E+00 | 1.1737373E+02 |
| 12 | 2.8941294E+00 | 1.3405239E+02 |
| 13 | 3.6649873E+00 | 1.0534641E+02 |
| 14 | 1.3627792E+00 | 1.2335949E+02 |
| 15 | 3.7187978E+00 | 1.2502106E+02 |
| 16 | 1.8658681E+00 | 1.1207542E+02 |
| 17 | 3.2728091E+00 | 1.1758804E+02 |
| 18 | 3.9175915E+00 | 1.0100987E+02 |

After execution of the code, the data are contained in a variable called `df`. This is a structure referred to as a Pandas *DataFrame*.

A **Pandas dataframe** is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it as a spreadsheet.

To see the contents of `df`, simply use:

```
1 df
```

| | | calcium | sodium |
|----|----|----------|------------|
| 1 | | | |
| 2 | 0 | 3.455582 | 112.690980 |
| 3 | 1 | 3.669026 | 125.663330 |
| 4 | 2 | 2.789910 | 105.821810 |
| 5 | 3 | 2.939900 | 98.172772 |
| 6 | 4 | 5.426060 | 97.931489 |
| 7 | 5 | 0.715811 | 120.858330 |
| 8 | 6 | 5.652390 | 112.871500 |
| 9 | 7 | 3.571320 | 112.647360 |
| 10 | 8 | 4.300067 | 132.031720 |
| 11 | 9 | 1.369419 | 118.499010 |
| 12 | 10 | 2.550962 | 117.373730 |
| 13 | 11 | 2.894129 | 134.052390 |
| 14 | 12 | 3.664987 | 105.346410 |
| 15 | 13 | 1.362779 | 123.359490 |
| 16 | 14 | 3.718798 | 125.021060 |
| 17 | 15 | 1.865868 | 112.075420 |
| 18 | 16 | 3.272809 | 117.588040 |
| 19 | 17 | 3.917591 | 101.009870 |

(Compare with the result of `print(df)` which displays the contents in a different format.)

The output shows in the first column an index, integers from 0 to 17; and the calcium and sodium concentrations in columns 2 and 3, respectively. The default indexing starts from zero (Python is a 'zero-based' programming language).

In a dataframe, the first column is referred to as *Indices*, the first row is referred to as *Labels*. Note that the row with the labels is excluded from the row count. Similarly, the row with the indices is excluded from the column count.

For large data sets, the function `head` is a convenient way to get a feel of the dataset.

```
1 df.head()
```

| | | calcium | sodium |
|---|---|----------|------------|
| 1 | | | |
| 2 | 0 | 3.455582 | 112.690980 |
| 3 | 1 | 3.669026 | 125.663330 |
| 4 | 2 | 2.789910 | 105.821810 |
| 5 | 3 | 2.939900 | 98.172772 |
| 6 | 4 | 5.426060 | 97.931489 |

Without any input argument, this displays the first five data lines of the dataframe. You can specify alter the number of rows displayed by including a single integer as argument, e.g. `head(10)`.

If you feel there are too many decimal places in the default view, you can restrict their number by using the `round` function:

```
1 df.head().round(2)
```

Data Frames

```
1      calcium  sodium
2  0      3.46  112.69
3  1      3.67  125.66
4  2      2.79  105.82
5  3      2.94   98.17
6  4      5.43   97.93
```

While we can see how many rows there are in a dataframe when we display the whole data frame and look at the last index, there is a convenient way to obtain the number directly:

```
1 no_rows = len(df)
2
3 print('Data frame has', no_rows, 'rows')
```

```
1 Data frame has 18 rows
```

You could see above, that the columns of the dataframe have labels. To see all labels:

```
1 column_labels = df.columns
2
3 print(column_labels)
```

```
1 Index(['calcium', 'sodium'], dtype='object')
```

Now we can count the labels to obtain the number of columns:

```
1 no_columns = len(column_labels)
2
3 print('Data frame has', no_columns, 'columns')
```

```
1 Data frame has 2 columns
```

And if you want to have both the number of the rows and the columns together, use `shape`. Shape returns a tuple of two numbers, first the number of rows, then the number of columns.

```
1 df_shape = df.shape
2
3 print('Data frame has', df_shape[0], 'rows and', df_shape[1], 'columns'
      )
```

```
1 Data frame has 18 rows and 2 columns
```

Notice that `shape` (like `columns`) is not followed by round parenthesis. It is not a function that can take arguments. Technically, `shape` is a 'property' of the dataframe.

To find out what data type is contained in each of the columns, us `dtypes`, another 'property':

```
1 df.dtypes
```

```
1 calcium    float64
2 sodium     float64
3 dtype: object
```

In this case, both columns contain floating point (decimal) numbers.

DIY1: Read data into a dataframe

Download the data file 'loan_data.csv' using the link given above in "Materials for this Lesson". It contains data that can be used for the assessment of loan applications. Read the data into a DataFrame. It is best to assign it a name other than 'df' (to avoid overwriting the Evereley data set).

Display the first ten rows of the Loan data set to see its contents. It is taken from a tutorial on Data Handling in Python which you might find useful for further practice.

From this exercise we can see that a dataframe can contain different types of data: real numbers (e.g. LoanAmount), integers (ApplicantIncome), categorical data (Gender), and strings (Loan_ID).

```
1 from pandas import read_csv
2
3 # dataframe from .csv file
4 df_loan = read_csv("data/loan_data.csv")
5
6 # display contents
7 df_loan.head(10)
```

| | Loan_ID | Gender | Married | ... | Loan_Amount_Term | Credit_History | |
|----|---------|-----------|---------|-----|------------------|----------------|-----|
| 2 | 0 | LP001015 | Male | Yes | ... | 360.0 | 1.0 |
| | | Urban | | | | | |
| 3 | 1 | LP001022 | Male | Yes | ... | 360.0 | 1.0 |
| | | Urban | | | | | |
| 4 | 2 | LP001031 | Male | Yes | ... | 360.0 | 1.0 |
| | | Urban | | | | | |
| 5 | 3 | LP001035 | Male | Yes | ... | 360.0 | NaN |
| | | Urban | | | | | |
| 6 | 4 | LP001051 | Male | No | ... | 360.0 | 1.0 |
| | | Urban | | | | | |
| 7 | 5 | LP001054 | Male | Yes | ... | 360.0 | 1.0 |
| | | Urban | | | | | |
| 8 | 6 | LP001055 | Female | No | ... | 360.0 | 1.0 |
| | | Semiurban | | | | | |
| 9 | 7 | LP001056 | Male | Yes | ... | 360.0 | 0.0 |
| | | Rural | | | | | |
| 10 | 8 | LP001059 | Male | Yes | ... | 240.0 | 1.0 |
| | | Urban | | | | | |

```
11 9  LP001067  Male  No  ...  360.0  1.0
12    Semiurban
13 [10 rows x 12 columns]
```

Accessing data in a DataFrame

If a datafile is large and you only want to check the format of data in a specific column, you can limit the display to that column. To access data contained in a specific column of a dataframe, we can use a similar convention as in a Python dictionary, treating the column names as 'keys'. E.g. to show all rows in column 'Calcium', use:

```
1 df['calcium']
```

```
1 0    3.455582
2 1    3.669026
3 2    2.789910
4 3    2.939900
5 4    5.426060
6 5    0.715811
7 6    5.652390
8 7    3.571320
9 8    4.300067
10 9    1.369419
11 10   2.550962
12 11   2.894129
13 12   3.664987
14 13   1.362779
15 14   3.718798
16 15   1.865868
17 16   3.272809
18 17   3.917591
19 Name: calcium, dtype: float64
```

To access individual rows of a column we use two pairs of square brackets:

```
1 df['calcium'][0:3]
```

```
1 0    3.455582
2 1    3.669026
3 2    2.789910
4 Name: calcium, dtype: float64
```

Here all rules for slicing can be applied. As for lists and tuples, the indexing of rows is semi-inclusive, lower boundary included, upper boundary excluded. Note that the first pair of square brackets refers to a column and the second pair refers to the rows. This is different from e.g. accessing items in a

nested list.

Accessing items in a Pandas dataframe is analogous to accessing the values in a Python dictionary by referring to its keys.

To access non-contiguous elements, we use an additional pair of square brackets (as if for a list within a list):

```
1 df['calcium'][[1, 3, 7]]
```

```
1 1    3.669026
2 3    2.939900
3 7    3.571320
4 Name: calcium, dtype: float64
```

Another possibility to index and slice a dataframe is the use of the ‘index location’ or `iloc` property. It refers first to rows and then to columns by index, all within a single pair of brackets. For example, to get all rows : of the first column (index 0), you use:

```
1 df.iloc[:, 0]
```

```
1 0    3.455582
2 1    3.669026
3 2    2.789910
4 3    2.939900
5 4    5.426060
6 5    0.715811
7 6    5.652390
8 7    3.571320
9 8    4.300067
10 9    1.369419
11 10   2.550962
12 11   2.894129
13 12   3.664987
14 13   1.362779
15 14   3.718798
16 15   1.865868
17 16   3.272809
18 17   3.917591
19 Name: calcium, dtype: float64
```

To display only the first three calcium concentrations, you use slicing, remembering that the upper bound is excluded):

```
1 df.iloc[0:3, 0]
```

```
1 0    3.455582
2 1    3.669026
3 2    2.789910
```

```
4 Name: calcium, dtype: float64
```

To access non-consecutive values, we can use a pair of square brackets within the pair of square brackets:

```
1 df.iloc[[2, 4, 7], 0]
```

```
1 2    2.78991
2 4    5.42606
3 7    3.57132
4 Name: calcium, dtype: float64
```

Similarly, we can access the values from multiple columns:

```
1 df.iloc[[2, 4, 7], :]
```

```
1      calcium      sodium
2 2    2.78991  105.821810
3 4    5.42606   97.931489
4 7    3.57132  112.647360
```

To pick only the even rows from the two columns, check this colon notation:

```
1 df.iloc[:18:2, :]
```

```
1      calcium      sodium
2 0    3.455582  112.690980
3 2    2.789910  105.821810
4 4    5.426060   97.931489
5 6    5.652390  112.871500
6 8    4.300067  132.031720
7 10   2.550962  117.373730
8 12   3.664987  105.346410
9 14   3.718798  125.021060
10 16   3.272809  117.588040
```

The number after the second colon indicates the stepsize.

DIY2: Select data from dataframe

Display the calcium and sodium concentrations of all patients except the first. Check the model solution at the bottom for options.

```
1 df[['calcium', 'sodium']][1:]
```

```
1      calcium      sodium
2 1    3.669026  125.663330
```

```
3 2 2.789910 105.821810
4 3 2.939900 98.172772
5 4 5.426060 97.931489
6 5 0.715811 120.858330
7 6 5.652390 112.871500
8 7 3.571320 112.647360
9 8 4.300067 132.031720
10 9 1.369419 118.499010
11 10 2.550962 117.373730
12 11 2.894129 134.052390
13 12 3.664987 105.346410
14 13 1.362779 123.359490
15 14 3.718798 125.021060
16 15 1.865868 112.075420
17 16 3.272809 117.588040
18 17 3.917591 101.009870
```

Mixing the ways to access specific data in a dataframe can be confusing and needs practice.

Search for missing values

Some tables contain missing entries. You can check a dataframe for such missing entries. If no missing entry is found, the function `isnull` will return `False`.

```
1 df.isnull().any()
```

```
1 calcium    False
2 sodium     False
3 dtype: bool
```

This shows that there are no missing entries in our dataframe.

DIY3: Find NaN in dataframe

In the Loan data set, check the entry 'Self-employed' for ID LP001059. It shows how a missing value is represented as 'NaN' (not a number).

Verify that the output of `isnull` in this case is `True`

```
1 df_loan['Self_Employed'][8]
```

```
1 nan
```

```
1 df_loan['Self_Employed'][8:9].isnull()
```

```
1 8      True
2 Name: Self_Employed, dtype: bool
```

Basic data features

Summary Statistics

To get a summary of basic data features use the function `describe`:

```
1 description = df.describe()
2
3 description
```

```
1          calcium      sodium
2 count  18.000000    18.000000
3 mean    3.174301    115.167484
4 std     1.306652     10.756852
5 min     0.715811     97.931489
6 25%     2.610699    107.385212
7 50%     3.364195    115.122615
8 75%     3.706355    122.734200
9 max     5.652390    134.052390
```

The `describe` function produces a new dataframe (here called 'description') that contains the number of samples, the mean, the standard deviation, minimum, 25th, 50th, 75th percentile, and the maximum value for each column of the data. Note that the indices of the rows have now been replaced by strings. To access rows, it is possible to refer to those names using the `loc` property. E.g. to access the mean of the calcium concentrations from the description, each of the following is valid:

```
1 # Option 1
2 description.loc['mean']['calcium']
3
4 # Option 2
```

```
1 3.1743005405555555
```

```
1 description.loc['mean'][0]
2
3 # Option 3
```

```
1 3.1743005405555555
```

```
1 description['calcium']['mean']
2
3 # Option 4
```

```
1 3.1743005405555555
```

```
1 description['calcium'][1]
```

```
1 3.1743005405555555
```

DIY4: Practice

Use your own .csv data set to practice. (If you don't have a data set at hand, any excel table can be exported as .csv.) Read it into a dataframe, check its header, access individual values or sets of values. Create a statistics using `describe` and check for missing values using `.isnull`.

[ad libitum]

Iterating through the columns

Now we know how to access all data in a dataframe and how to get a summary statistics over each column.

Here is code to iterate through the columns and access the first two concentrations:

```
1 for col in df:
2
3     print(df[col][0:2])
```

```
1 0    3.455582
2 1    3.669026
3 Name: calcium, dtype: float64
4 0    112.69098
5 1    125.66333
6 Name: sodium, dtype: float64
```

As a slightly more complex example, we access the median ('50%') of each column in the description and add it to a list:

```
1 conc_medians = list()
2
3 for col in df:
4
5     conc_medians.append(df[col].describe()['50%'])
6
7 print('The columns are: ', list(df.columns))
```

```
1 The columns are:  ['calcium', 'sodium']
```

```
1 print('The medians are: ', conc_medians)
```

```
1 The medians are: [3.3641954, 115.122615]
```

This approach is useful for data frames with a large number of columns. For instance, it is possible to then create a boxplot or histogram for the means, medians etc. of the dataframe and thus to get an overview of all (comparable) columns.

Selecting a subset based on a template

An analysis of a data set may need to be done on part of the data. This can often be formulated by using a logical condition which specifies the required subset.

For this we will assume that some of the data are labelled '0' and some are labelled '1'. Let us therefore see how to add a new column to our Evereleys data frame which contains the (in this case arbitrary) labels.

First we randomly create as many labels as we have rows in the data frame. We can use the `randint` function which we import from 'numpy.random'. `randint` in its simple form takes two arguments. First the upper bound of the integer needed, where by default it starts from zero. As Python is exclusive on the upper bound, providing '2' will thus yield either '0' or '1' only.

```
1 from numpy.random import randint
2
3 no_rows = len(df)
4 randomLabel = randint(2, size=no_rows)
5
6 print('Number of rows: ', no_rows)
```

```
1 Number of rows: 18
```

```
1 print('Number of Labels:', len(randomLabel))
```

```
1 Number of Labels: 18
```

```
1 print('Labels: ', randomLabel)
```

```
1 Labels: [0 1 0 0 0 0 0 1 1 0 1 1 0 1 0 1 0 1]
```

Note how we obtain the number of rows (18) using `len` and do not put it directly into the code.

Now we create a new data column in our `df` dataframe which contains the labels. To create a new column, you can simply refer to a column name that does not yet exist and assign values to it. Let us call it 'gender', assuming that '0' represents male and '1' represents female.

As gender specification can include more than two labels, try to create a column with more than two randomly assigned labels e.g. (0, 1, 2).

```
1 df['gender'] = randomLabel
2
3 df.head()
```

| | | calcium | sodium | gender |
|---|---|----------|------------|--------|
| 2 | 0 | 3.455582 | 112.690980 | 0 |
| 3 | 1 | 3.669026 | 125.663330 | 1 |
| 4 | 2 | 2.789910 | 105.821810 | 0 |
| 5 | 3 | 2.939900 | 98.172772 | 0 |
| 6 | 4 | 5.426060 | 97.931489 | 0 |

Now we can use the information contained in 'gender' to filter the data by gender. To achieve this, we use a conditional statement. Let us check which of the rows are labelled as '1':

```
1 df['gender'] == 1
```

```
1 0    False
2 1     True
3 2    False
4 3    False
5 4    False
6 5    False
7 6    False
8 7     True
9 8     True
10 9    False
11 10    True
12 11    True
13 12    False
14 13    True
15 14    False
16 15    True
17 16    False
18 17    True
19 Name: gender, dtype: bool
```

If we assign the result of the conditional statement (Boolean True or False) to a variable, then this variable can act as a template to filter the data. If we call the data frame with that variable, we will only get the rows where the condition was found to be True:

```
1 df_female = df['gender'] == 1
2
3 df[df_female]
```

| | | calcium | sodium | gender |
|---|---|----------|-----------|--------|
| 2 | 1 | 3.669026 | 125.66333 | 1 |

| | | | | |
|---|----|----------|-----------|---|
| 3 | 7 | 3.571320 | 112.64736 | 1 |
| 4 | 8 | 4.300067 | 132.03172 | 1 |
| 5 | 10 | 2.550962 | 117.37373 | 1 |
| 6 | 11 | 2.894129 | 134.05239 | 1 |
| 7 | 13 | 1.362779 | 123.35949 | 1 |
| 8 | 15 | 1.865868 | 112.07542 | 1 |
| 9 | 17 | 3.917591 | 101.00987 | 1 |

Using the Boolean, we only pick the rows that are labelled '1' and thus get a subset of the data according to the label.

DIY5: Using a template

Modify the code to calculate the number of samples labelled 0 and check the number of rows of that subset.

```
1 from numpy.random import randint
2
3 no_rows = len(df)
4
5 randomLabel = randint(2, size=no_rows)
6
7 df['gender'] = randomLabel
8
9 df_male = df[df['gender'] == 0]
10
11 no_males = len(df[df_male])
12
13 print(no_males, 'samples are labelled "male".')
```

```
1 10 samples are labelled "male".
```

Visualisation of data

It is easy to see from the numbers that the concentrations of sodium are much higher than that of calcium. However, to also compare the medians, percentiles and the spread of the data it is better to use visualisation.

The simplest way of visualisation is to use Pandas functionality which offers direct ways of plotting. Here is an example where a boxplot is created for each column:

```
1 import matplotlib.pyplot as plt
```

```
1 Matplotlib is building the font cache; this may take a moment.
```



```
1 df = read_csv("data/everleys_data.csv")
2 plt.boxplot(df)
```

```
1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f1271abdf30>, <
  matplotlib.lines.Line2D object at 0x7f1271abe200>, <matplotlib.lines
  .Line2D object at 0x7f1271abf310>, <matplotlib.lines.Line2D object
  at 0x7f1271abf5e0>], 'caps': [<matplotlib.lines.Line2D object at 0
  x7f1271abe590>, <matplotlib.lines.Line2D object at 0x7f1271abe7a0>,
  <matplotlib.lines.Line2D object at 0x7f1271abf8b0>, <matplotlib.
  lines.Line2D object at 0x7f1271abfb80>], 'boxes': [<matplotlib.lines
  .Line2D object at 0x7f1271abdc60>, <matplotlib.lines.Line2D object
  at 0x7f1271abf040>], 'medians': [<matplotlib.lines.Line2D object at
  0x7f1271abea70>, <matplotlib.lines.Line2D object at 0x7f1271abfe50
  >], 'fliers': [<matplotlib.lines.Line2D object at 0x7f1271abed40>, <
  matplotlib.lines.Line2D object at 0x7f1271b04160>], 'means': []}
```

```
1 plt.show()
```

By default, boxplots are shown for all columns if no further argument is given to the function (empty round parenthesis). As the calcium plot is rather squeezed we may wish to see it individually. This can be done by specifying the calcium column as an argument:

```
1 # Boxplot of calcium results
2 # df.boxplot(column='calcium');
3
4
5 plt.boxplot(df['calcium'])
```

```
1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f1271ba1e40>, <
  matplotlib.lines.Line2D object at 0x7f1271ba1ff0>], 'caps': [<
  matplotlib.lines.Line2D object at 0x7f1271ba22c0>, <matplotlib.lines
  .Line2D object at 0x7f1271ba2590>], 'boxes': [<matplotlib.lines.
  Line2D object at 0x7f1271ba1b70>], 'medians': [<matplotlib.lines.
  Line2D object at 0x7f1271ba2860>], 'fliers': [<matplotlib.lines.
  Line2D object at 0x7f1271ba2b30>], 'means': []}
```

```
1 plt.show()
```

Plots using Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

The above is an easy way to create boxplots directly on the dataframe. It is based on the library Matplotlib and specifically uses the **pyplot library**. For simplicity, the code is put in a convenient Pandas function.

However, we are going to use **Matplotlib** extensively later on in the course, and we therefore now introduce the direct, generic way of using it.

For this, we import the function `subplots` from the `pyplot` library:

```
1 from matplotlib.pyplot import subplots
```

The way to use `subplots` is to first set up a figure environment (below it is called 'fig') and an empty coordinate system (below called 'ax'). The plot is then done using one of the many methods available in Matplotlib. We apply it to the coordinate system 'ax'.

As an example, let us create a boxplot of the calcium variable. As an argument of the function we need to specify the data. We can use the values of the 'calcium' concentrations from the column with that name:

```
1 fig, ax = subplots()
2
3 ax.boxplot(df['calcium'])
```

```
1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f12719fed0>, <
  matplotlib.lines.Line2D object at 0x7f12719ff070>], 'caps': [<
  matplotlib.lines.Line2D object at 0x7f12719ff340>, <matplotlib.lines.
  .Line2D object at 0x7f12719ff610>], 'boxes': [<matplotlib.lines.
  Line2D object at 0x7f12719fe830>], 'medians': [<matplotlib.lines.
  Line2D object at 0x7f12719ff8e0>], 'fliers': [<matplotlib.lines.
  Line2D object at 0x7f12719ffbb0>], 'means': []}
```

```
1 ax.set_title('Boxplot of Everley\'s Calcium');
```

Note how following the actual plot we define the title of the plot by referring to the same coordinate system `ax`.

The value of `subplots` becomes apparent when we try to create more than one plot in a single figure.

Here is an example to create two boxplots next to each other. The keyword arguments to use is 'ncols' which is the number of figures per row. 'ncols=2' indicates that you want to have two plots next to each other.

```
1 fig, ax = subplots(ncols=2)
2
3 ax[0].boxplot(df['calcium'])
```

```
1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f1271a58040>, <
  matplotlib.lines.Line2D object at 0x7f1271a58310>], 'caps': [<
  matplotlib.lines.Line2D object at 0x7f1271a585e0>, <matplotlib.lines.
  .Line2D object at 0x7f1271a58820>], 'boxes': [<matplotlib.lines.
  Line2D object at 0x7f1271a33a90>], 'medians': [<matplotlib.lines.
```



Figure 1: png

```
Line2D object at 0x7f1271a58af0>], 'fliers': [<matplotlib.lines.
Line2D object at 0x7f1271a58dc0>], 'means': []}
```

```
1 ax[0].set_title('Calcium')
2
3 ax[1].boxplot(df['sodium'])
```

```
1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f1271a59600>, <
matplotlib.lines.Line2D object at 0x7f1271a598d0>], 'caps': [<
matplotlib.lines.Line2D object at 0x7f1271a59ba0>, <matplotlib.lines
matplotlib.lines.Line2D object at 0x7f1271a59e70>], 'boxes': [<matplotlib.lines.
matplotlib.lines.Line2D object at 0x7f1271a59330>], 'medians': [<matplotlib.lines.
matplotlib.lines.Line2D object at 0x7f1271a5a140>], 'fliers': [<matplotlib.lines.
matplotlib.lines.Line2D object at 0x7f1271a5a410>], 'means': []}
```

```
1 ax[1].set_title('Sodium');
```

Note that you now have to refer to each of the subplots by indexing the coordinate system 'ax'. | This figure gives a good overview of the Everley's data.

If you prefer to have the boxplots of both columns in a single figure, that can also be done:

```
1 fig, ax = subplots(ncols=1, nrows=1)
2
```



Figure 2: png

```
3 ax.boxplot([df['calcium'], df['sodium']], positions=[1, 2])
```

```
1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f12718b8790>, <matplotlib.lines.Line2D object at 0x7f12718b8e80>, <matplotlib.lines.Line2D object at 0x7f12718b9f60>, <matplotlib.lines.Line2D object at 0x7f12718ba230>], 'caps': [<matplotlib.lines.Line2D object at 0x7f12718b9150>, <matplotlib.lines.Line2D object at 0x7f12718b9420>, <matplotlib.lines.Line2D object at 0x7f12718ba500>, <matplotlib.lines.Line2D object at 0x7f12718ba7d0>], 'boxes': [<matplotlib.lines.Line2D object at 0x7f12718b8ac0>, <matplotlib.lines.Line2D object at 0x7f12718b9c90>], 'medians': [<matplotlib.lines.Line2D object at 0x7f12718b96f0>, <matplotlib.lines.Line2D object at 0x7f12718baaa0>], 'fliers': [<matplotlib.lines.Line2D object at 0x7f12718b99c0>, <matplotlib.lines.Line2D object at 0x7f12718bad70>], 'means': []}
```

```
1 ax.set_title('Boxplot of Calcium (left) and Sodium (right)')
```

DIY6: Boxplot from Loan data

Plot the boxplots of the 'ApplicantIncome' and the 'CoapplicantIncome' in the Loan data using the above code.

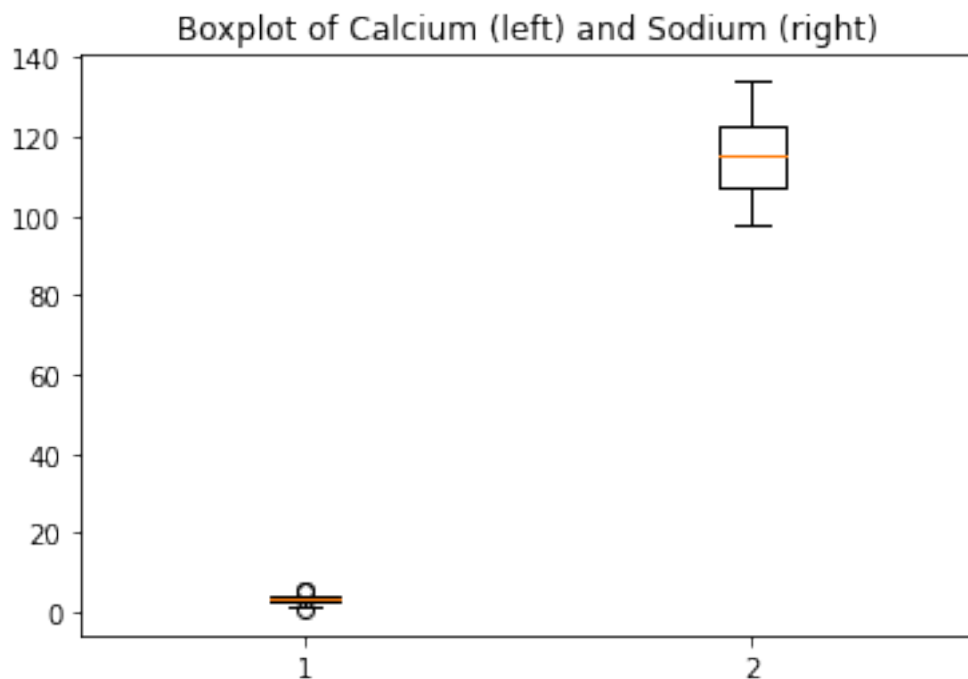


Figure 3: png

```

1 fig, ax = subplots(ncols=1, nrows=1)
2
3 ax.boxplot([df_loan['ApplicantIncome'], df_loan['CoapplicantIncome']],
              positions=[1, 2])

```

```

1 {'whiskers': [<matplotlib.lines.Line2D object at 0x7f127192c2b0>, <
  matplotlib.lines.Line2D object at 0x7f127192c580>, <matplotlib.lines
  .Line2D object at 0x7f127192d660>, <matplotlib.lines.Line2D object
  at 0x7f127192d930>], 'caps': [<matplotlib.lines.Line2D object at 0
  x7f127192c850>, <matplotlib.lines.Line2D object at 0x7f127192cb20>,
  <matplotlib.lines.Line2D object at 0x7f127192dc00>, <matplotlib.
  lines.Line2D object at 0x7f127192ded0>], 'boxes': [<matplotlib.lines
  .Line2D object at 0x7f12718f0df0>, <matplotlib.lines.Line2D object
  at 0x7f127192d390>], 'medians': [<matplotlib.lines.Line2D object at
  0x7f127192cdf0>, <matplotlib.lines.Line2D object at 0x7f127192e1a0
  >], 'fliers': [<matplotlib.lines.Line2D object at 0x7f127192d0c0>, <
  matplotlib.lines.Line2D object at 0x7f127192e470>], 'means': []}

```

```

1 ax.set_title('Applicant Income (left) & Co-Applicant Income (right)');

```

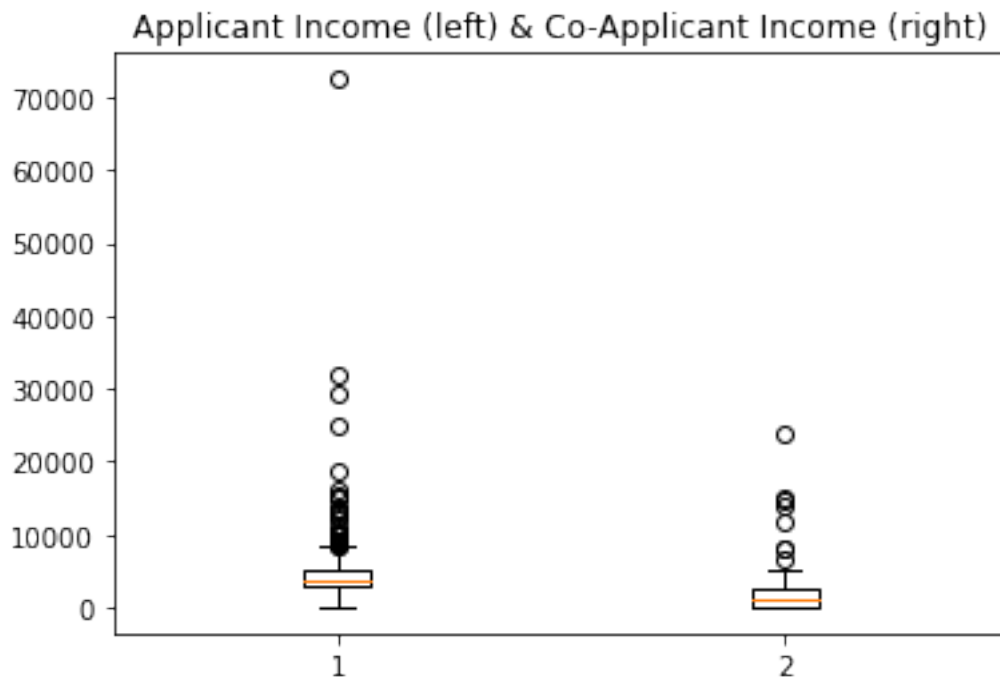


Figure 4: Applicant Income (left) & Co-Applicant Income (right)

Histogram

Another good overview is the histogram: Containers or ‘bins’ are created over the range of values found within a column and the count of the values for each bin is plotted on the vertical axis.

```
1 fig, ax = subplots(ncols=2, nrows=1)
2
3 ax[0].hist(df['calcium'])
```

```
1 (array([1., 2., 1., 1., 3., 5., 2., 1., 0., 2.]), array([0.71581063,
2          1.20946859, 1.70312654, 2.1967845 , 2.69044246,
3          3.18410042, 3.67775837, 4.17141633, 4.66507429, 5.15873224,
          5.6523902 ]), <BarContainer object of 10 artists>)
```

```
1 ax[0].set(xlabel='Calcium', ylabel='Count');
2
3 ax[1].hist(df['sodium'])
```

```
1 (array([3., 0., 2., 1., 3., 3., 1., 3., 0., 2.]), array([ 97.931489 ,
2          101.5435791, 105.1556692, 108.7677593, 112.3798494,
3          115.9919395, 119.6040296, 123.2161197, 126.8282098, 130.4402999,
          134.05239 ]), <BarContainer object of 10 artists>)
```

```

1 ax[1].set(xlabel='Sodium', ylabel='Count');
2
3 fig.suptitle('Histograms of Everley concentrations', fontsize=15);

```

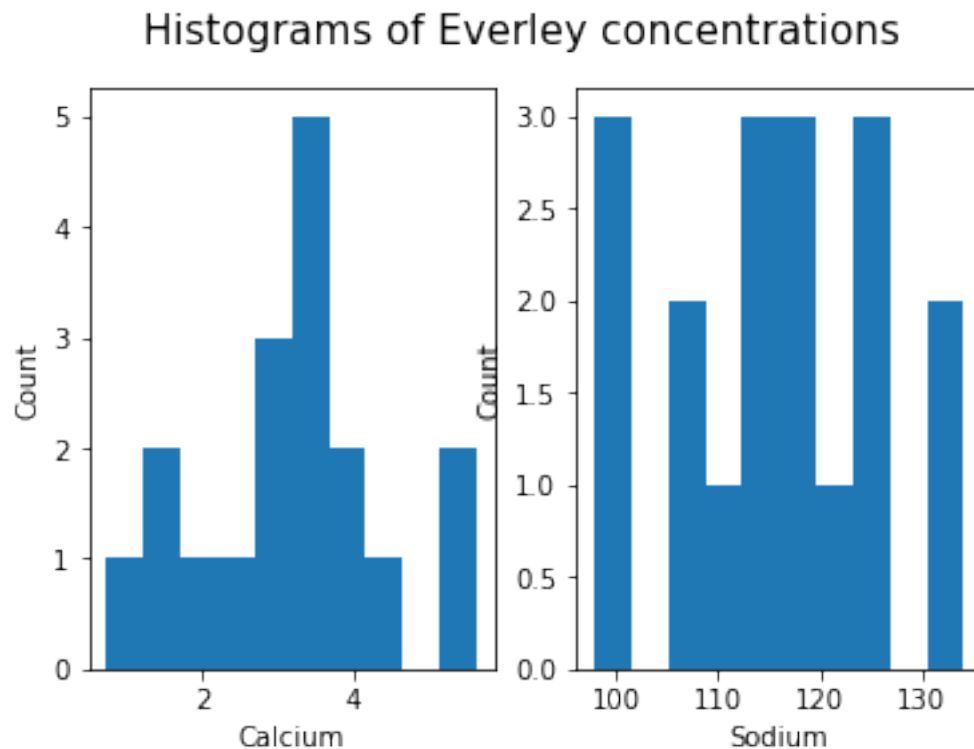


Figure 5: png

This also shows how to add labels to the axes and a title to the overall figure.

This uses the default value for the generation of the bins. It is set to 10 bins over the range of which values are found. The number of bins in the histogram can be changed using the keyword argument 'bins':

```

1 fig, ax = subplots(ncols=2, nrows=1)
2
3 ax[0].hist(df['calcium'], bins=5)

```

```

1 (array([3., 2., 8., 3., 2.]), array([0.71581063, 1.70312654,
2     2.69044246, 3.67775837, 4.66507429,
3     5.6523902 ]), <BarContainer object of 5 artists>)

```

```

1 ax[0].set(xlabel='Calcium, 5 bins', ylabel='Count');
2
3 ax[1].hist(df['calcium'], bins=15)

```

```

1 (array([1., 2., 0., 1., 0., 1., 3., 1., 4., 2., 1., 0., 0., 0., 2.]),
   array([0.71581063, 1.04491593, 1.37402124, 1.70312654, 2.03223185,
2         2.36133715, 2.69044246, 3.01954776, 3.34865307, 3.67775837,
3         4.00686368, 4.33596898, 4.66507429, 4.99417959, 5.3232849 ,
4         5.6523902 ]), <BarContainer object of 15 artists>)

```

```

1 ax[1].set(xlabel='Calcium, 15 bins', ylabel='Count');
2 fig.suptitle('Histograms with Different Binnings', fontsize=16);

```

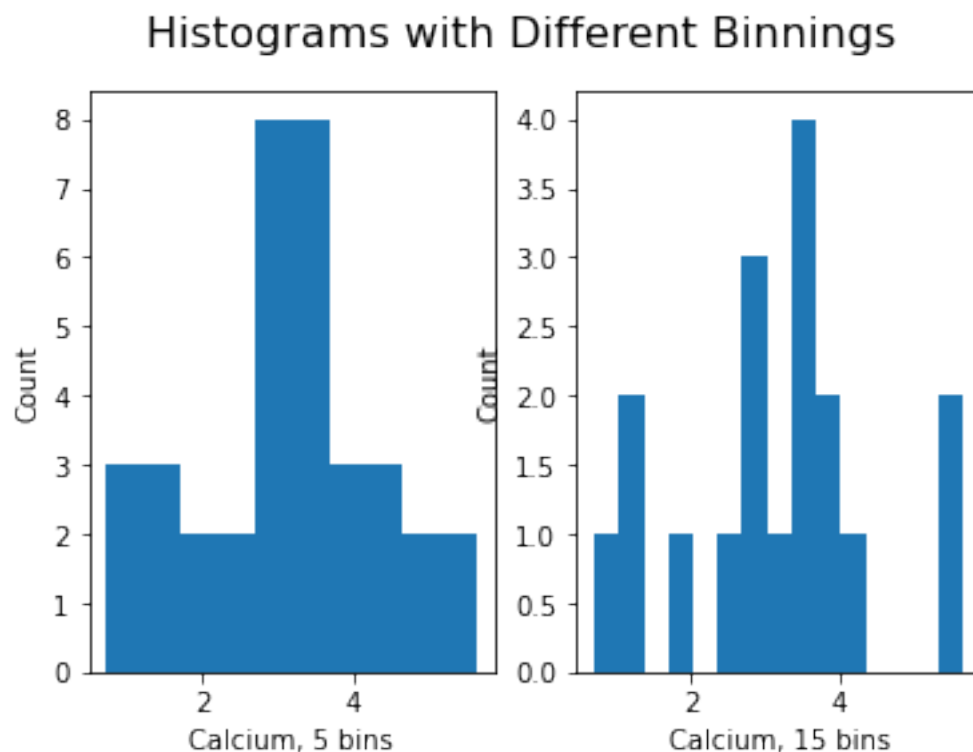


Figure 6: png

Note how the y-label of the right figure is not placed well. To correct for the placement of the labels and the title, you can use `tight_layout` on the figure:

```

1 fig, ax = subplots(ncols=2, nrows=1)
2
3 ax[0].hist(df['calcium'], bins=5)

```

```

1 (array([3., 2., 8., 3., 2.]), array([0.71581063, 1.70312654,
2         2.69044246, 3.67775837, 4.66507429,
3         5.6523902 ]), <BarContainer object of 5 artists>)

```

```

1 ax[0].set(xlabel='Calcium, 5 bins', ylabel='Count');

```



```
2
3 ax[1].hist(df['calcium'], bins=15)
```

```
1 (array([1., 2., 0., 1., 0., 1., 3., 1., 4., 2., 1., 0., 0., 0., 2.]),
   array([0.71581063, 1.04491593, 1.37402124, 1.70312654, 2.03223185,
          2.36133715, 2.69044246, 3.01954776, 3.34865307, 3.67775837,
          4.00686368, 4.33596898, 4.66507429, 4.99417959, 5.3232849 ,
          5.6523902 ]), <BarContainer object of 15 artists>)
```

```
1 ax[1].set(xlabel='Calcium, 15 bins', ylabel='Count');
2 fig.suptitle('Histograms with Different Binnings', fontsize=16);
3 fig.tight_layout()
```

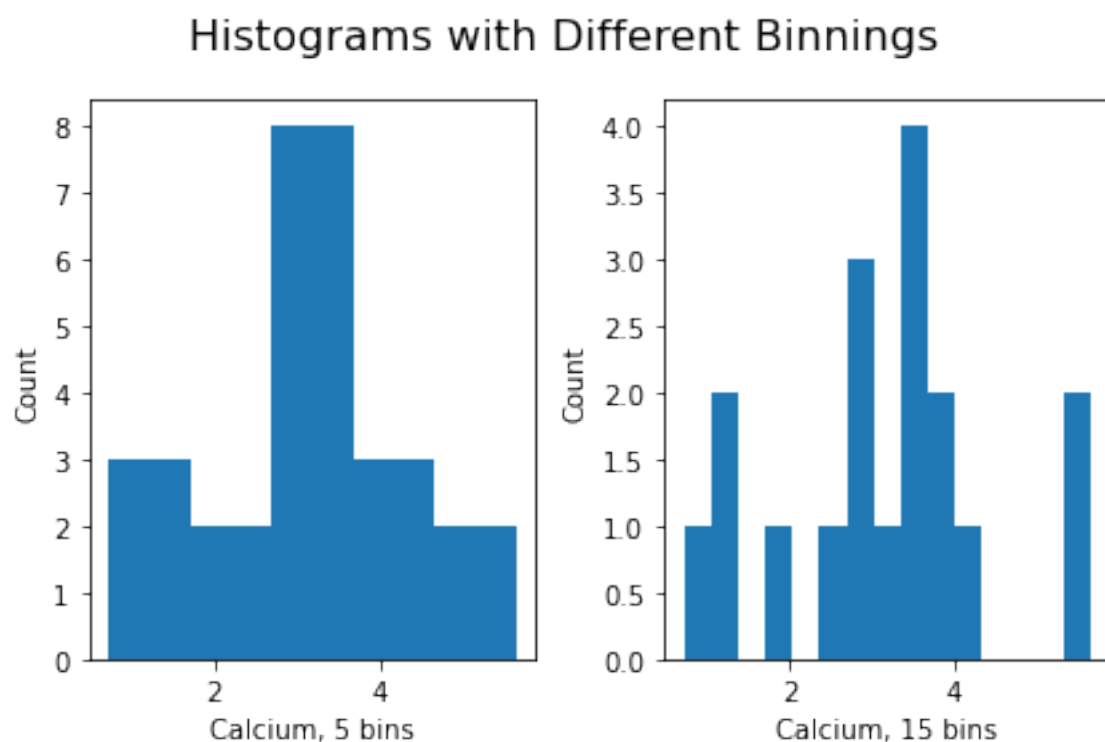


Figure 7: png

DIY7: Create the histogram of a column

Take the loan data and display the histogram of the loan amount that people asked for. (Loan amounts are divided by 1000, i.e. in k£ on the horizontal axis). Use e.g. 20 bins.

```
1 # Histogram of loan amounts in k£
2
3 fig, ax = subplots()
```

```
4
5 ax.hist(df_loan['LoanAmount'], bins=20)
```

```
1 (array([12., 34., 63., 97., 67., 33., 28., 9., 6., 2., 5., 1.,
2       1., 1., 1., 0., 1., 0., 0., 1.]), array([ 28. ,  54.1,  80.2,
3       106.3, 132.4, 158.5, 184.6, 210.7, 236.8,
4       262.9, 289. , 315.1, 341.2, 367.3, 393.4, 419.5, 445.6, 471.7,
       497.8, 523.9, 550. ]), <BarContainer object of 20 artists>)
```

```
1 ax.set(xlabel='Loan amount', ylabel='Count');
2
3 ax.set_title('Histograms of Loan Amounts', fontsize=16);
```

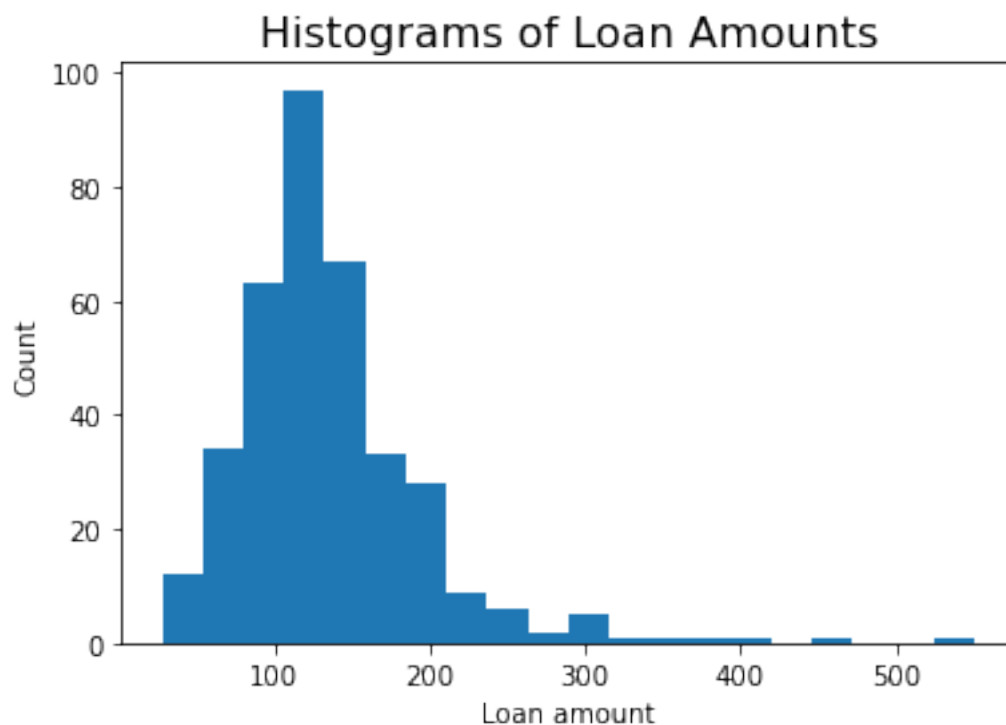


Figure 8: Histograms of Loan Amounts

Handling the Diabetes Data Set

We now return to the data set that started our enquiry into the handling of data in a dataframe.

We will now:

- Import the diabetes data from 'sklearn'

- Check the shape of the dataframe and search for NaNs
- Get a summary plot of one of its statistical quantities (e.g. mean) for all columns

First we import the data set and check its `head`. Wait until the numbers show below the code, it might take a while.

```
1 from sklearn import datasets
2
3 diabetes = datasets.load_diabetes()
4
5 X = diabetes.data
6
7 from pandas import DataFrame
8
9 df_diabetes = DataFrame(data=X)
10
11 df_diabetes.head()
```

```
1      0      1      2  ...      7      8      9
2  0  0.038076  0.050680  0.061696  ... -0.002592  0.019908 -0.017646
3  1 -0.001882 -0.044642 -0.051474  ... -0.039493 -0.068330 -0.092204
4  2  0.085299  0.050680  0.044451  ... -0.002592  0.002864 -0.025930
5  3 -0.089063 -0.044642 -0.011595  ...  0.034309  0.022692 -0.009362
6  4  0.005383 -0.044642 -0.036385  ... -0.002592 -0.031991 -0.046641
7
8 [5 rows x 10 columns]
```

If you don't see all columns, use the cursor to scroll to the right.

Now let's check the number of columns and rows.

```
1 no_rows = len(df_diabetes)
2 no_cols = len(df_diabetes.columns)
3
4 print('Rows:', no_rows, 'Columns:', no_cols)
```

```
1 Rows: 442 Columns: 10
```

There are 442 rows organised in 10 columns.

To get an overview, let us extract the mean of each column using 'describe' and plot all means as a bar chart. The Matplotlib function to plot a bar chart is `bar`:

```
1 conc_means = list()
2
3 for col in df_diabetes:
4     conc_means.append(df_diabetes[col].describe()['mean'])
5
6 print('The columns are: ', list(df_diabetes.columns))
```

```
1 The columns are: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
```

```
1 print('The medians are: ', conc_means, 2)
```

```
1 The medians are: [-3.6396225400041895e-16, 1.309912460049817e-16,
-8.013951493363262e-16, 1.2898179256674614e-16, -9.042540472060098e
-17, 1.3011211012575365e-16, -4.563971121592555e-16,
3.8631742350078977e-16, -3.848103334221131e-16, -3.39848812741592e
-16] 2
```

```
1 fig, ax = subplots()
2
3 bins = range(10)
4
5 ax.bar(bins, conc_means);
```

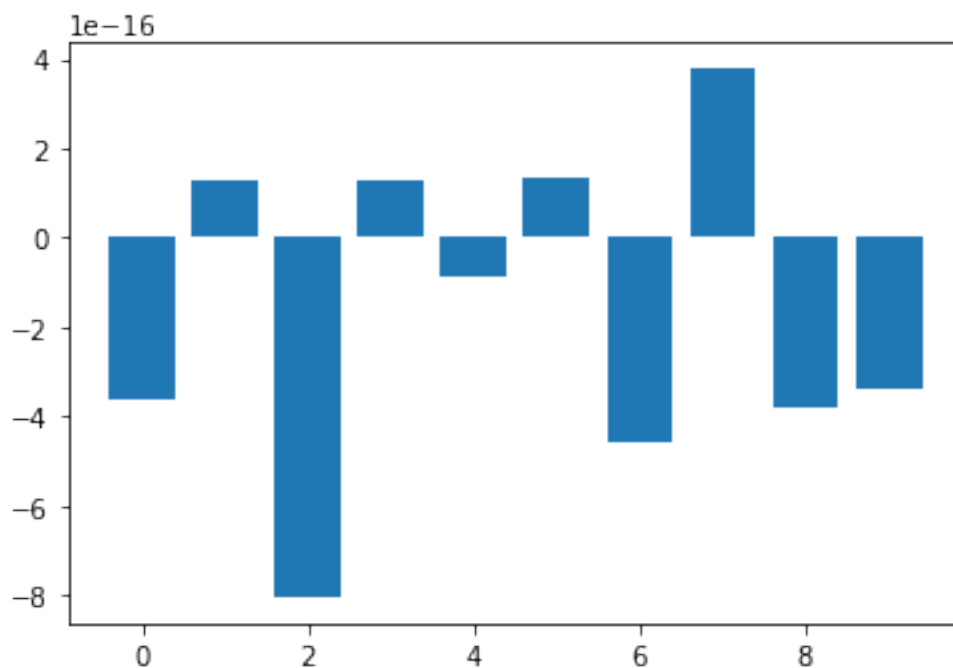


Figure 9: png

The bars in this plot go up and down. Note, however, that the vertical axis has values ranging from $-10^{(-15)}$ to $+10^{(-15)}$. This means that for all practical purposes all means are zero. This is not a coincidence. The original values have been normalised to zero mean for the purpose of applying some machine learning algorithm to them.

In this example, we see how important it is to check the data before working with them.

Assignment: The cervical cancer data set

Download the cervical cancer data set provided, import it using `read_csv`.

- How many rows and columns are there?
- How many columns contain floating point numbers (float64)?
- How many of the subjects are smokers?
- Calculate the percentage of smokers
- Plot the age distribution (with e.g. 50 bins)
- Get the mean and standard distribution of age of first sexual intercourse

Keypoints

- learned importing data set as dataframe
- learned inspecting data frame and accessing data
- Producing an overview of data features
- Creating data plots using matplotlib

Download the pdf

This section's pdf can be accessed at: [Download pdf](#)

Download jupyter notebook

This section's material in jupyter notebook can be accessed at: [Download notebook \(ipynb\)](#)