**Python**

|  |  |
| --- | --- |
| **Easter egg** | Import this |
| **Import pandas** | import pandas as pd |
| **Import NumPy** | import numpy as np |
| **Series (a column)** | pd.Series([21, 22, 23, 24], name = ‘age’)  *age = column name* |
| **Select an element (row; from 0 to …)** | table\_name[3]  *Select the 4th row* |
| **Make a custom axis to a Series** |  |
|  |  |
| **Index** is also called an **axis**; each element is called axis label.  Data in columns is called as **values**.  A series ideally should have the same datatype throughout its values (same format for the whole column). |  |
| **Dataframe** | A series is a column and a dataframe has multiple columns |
| **Create a dataframe** |  |
| **Columns and rows in a dataframe (table)** | df.shape  (3, 7)  3 rows and 7 columns |
| **Axis 0**  **Axis 1** | Vertical axis (rows)  Horizontal axis (column names) |
| **Select a column** | df[‘ADDRESS’]  Returns a series |
| **Select several columns** | df[[‘ADDRESS’, ‘CITY’]] |
| **Retrieve axis 1 / column names / information** | df.columns  Index([‘first\_name’, ‘last\_name’, ‘email’], dtype=‘object’) |
| **Retrieve axis 0 / vertical axis information** | df.index  RangeIndex(start=0, stop=3, step=1) or min 0, max 3 (3 rows) |
| **Retrieve a row** |  |
| **Make a column as an index** | E.g. you want to filter by last name. Instead of the default 0, 1, 2 etc. you can make the index as Nield, Scala Morrison.  *df.set\_index(‘column\_name’, inplace = True*) inplace is true so it edits the existing df instead of creating a new one  *df.loc[‘Scala’]* you can now use the last name to search for this row  **df.reset\_index(inplace = True)** reset the axis to default |
| **Copy a dataframe** | df2 = df.copy() |

**Importing Data**





[**Import CSV / read csv / load csv**](https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html)

Read csv as text:



Read csv into as a pandas dataframe:



**Read csv alternative:**

df = pd.read\_csv('https://raw.githubusercontent.com/thomasnield/machine-learning-demo-data/master/regression/winequality-red.csv')

df

**With *header* and *names*:**



**Alternative:**

Import pandas as pd

url = 'https://raw.githubusercontent.com/thomasnield/machine-learning-demo-data/master/timeseries/datetime\_formatting.csv'

df = pd.read\_csv(url)

df

**Format dates columns as dates while importing csv:**



**SQL** [**pd.read\_sql**](https://pandas.pydata.org/docs/reference/api/pandas.read_sql.html)

Import an SQL database:



Alternative:



If you don’t *parse\_dates*:

With parsing and without:



**JSON**

Read a JSON file:



Read a JSON file as a pandas dataframe:



**Selecting Rows and Columns**

**loc and iloc**

* **loc** works on **labels** assigned to the axis, **iloc** works on **numbers**.
* E.g. if axis is surname then **loc** will work only on the surname like ‘**vanli’** while **iloc** will only work on **numbers**.

**Select** (first two) **rows** df.iloc[0:2] or df.iloc[:2]

**Exclude first row** df.iloc[1:]

**Select all** df.iloc[:]

**All rows and columns 2-3** df.iloc[:, 1:3] you are selecting columns 2 and 3 with indexes 1 and 2

**Select last two columns** df.iloc[:, -2:] count from 0 to -1, -2 etc. from right to left. Select the column you want to have and with : you will select everything to the right

**Select the last row** df.iloc[-1]

**Select all rows and last column** df.iloc[:, -1]

**Select all rows and all columns except for the last column** df.iloc[:, :-1]

If you want to get the first 2 rows you need to select the third index, in this case 2:



**Select specific columns and rows**. In this case you will select column index 0 and column index 2 (first\_name and email). In the second query you will select second row and third column



**Select rows and columns using loc** df.loc[["samiam","thomasnield"], "email"]

**Reset index** df.reset\_index(inplace = True)

**Select values that start with a specific letter** condition = df["username"].str.startswith("s") username = column, s = letter start

df[condition]

**Multiple conditions. AND & OR are & and |** condition = df["username"].str.startswith("s") & df["email"].str.contains("gmail")

df[condition]

**Drop columns and rows**





**Remove columns:**



**Remove columns by selecting specific columns like column 1 and 4:**



**Adding rows and columns (appending)**

**Add a column at the end of the dataframe:**



**Add a column at a specific place:**



**Add a row:**



**Using concat (merge two datasets):**



**Updating data**

**Making a column in caps lock (upper):**



**Updating on a condition:**



**Update row on condition:**



**Unpivoting data (melting):**



Id\_vars = untouched columns

Value\_vars = columns that will be unpivoted

Var\_name = column names will be moved to this column

Value\_name = column values will be moved to this column



**Sorting, Casting, and Categories**

**Datatypes:**



**Timestamp example (date)** pd.Timestamp(‘20230130’) → 2023-01-30

**Difference between days** pd.Timestamp('20230130') - pd.Timestamp('20230127') → 3 days

**View datatypes** df.dtypes



You will get column name on the left (float, int, datatime etc.) and datatype on the right.

**Change a column to a different datatype** df[‘columnname’] = df[‘columnaname’].astype(‘bool’)

**Sort by 2 columns (first lightning, then rain inches)** df.sort\_values(by=["lightning","rain\_inches"])

df.sort\_values(by=["lightning","rain\_inches"],ascending=[False,True])

df.sort\_values(axis=0, ascending = True) *ascending if one value*

When using the sort methods, remember to **add the inplace=True parameter if you want to replace the existing dataframe** with the sorted one.

Sort by an row index or columns index df.sort\_index(axis = 0) for rows

df.sort\_index(axis = 1) for columns

**Replace a column with a *category* datatype:**

cat\_type = pd.CategoricalDtype(categories=["CLEAR", "MINOR", "MAJOR", "SEVERE"], ordered=True)

df["severity"] = df["severity"].astype(cat\_type) severity = column you want to replace

df.sort\_values(by=["severity"])

If you apply a categorization on a column that has values not mapping to any category, then those will become NA values.

**Python if/elif/else category function:**



**Apply this function to the wind\_speed\_mph column:**



**Add a new column:**



Categorize the last column and sort the data by that column in DESC:



**Removing Duplicative and Sparse Data**

**df used:**



**Get duplicated rows using .duplicated() function:**  df.duplicated()



It will mark rows that are duplicates.

If you want to see original rows and their duplicates (like in Excel) then add (keep=False):



**Look for duplicates in a specific column using *subset*:**



Or if you want to use multiple columns:



**Delete duplicates** df.drop\_duplicates(inplace=True)

**Delete duplicates based on a column** df.drop\_duplicates(subset=['record\_id'], inplace=True)

Number of unique values in a column df.nunique()



**Identify columns with single-values (e.g. value Shop in the whole column):**



**Drop these columns:** df.drop(delete\_cols, axis=1, inplace=True)

**Read csv file (open csv file) (using a link):**

wine\_df = pd.read\_csv('https://raw.githubusercontent.com/thomasnield/machine-learning-demo-data/master/regression/winequality-red.csv')

wine\_df

**Get number of rows and columns from .shape function:**



**Alternatively. Get number of rows ([0]) and columns ([1]) from the .shape function (where X = df):**



**Count the number of unique values per column:**



**Remove columns with 5% or less unique values:**



**Alternative using scikit-learn, VarianceThreshold and fit\_transform():**



Return from ndarray to get the columns using get\_support:



**Remove columns with duplicates and 3 or less unique values:**



**f-string / f’’ / f”” / printing with f / print(f:**



Example



**Handling missing data**

**Looking for missing values**

**Looking for missing values per row** df.isna() *not efficient in my opinion, alternatives:* df.notna(), df.isnull(), df.notnull()



**Check columns for missing values** df.isna().any() *or you can specify axis*  any(axis=0)



**Check rows for missing values** df.isna().any(axis = 1)



**Select (show) columns with NaN (null) values** df.loc[:, df.isna().any()]



**Select (show) rows with NaN (null) values** df.loc[df.isna().any(axis=1), :] *or* df.loc[df.isna().any(axis=1)]



**Look for missing values in specific columns** df[df['TEMPERATURE'].isna() | df['RAIN'].isna()]



**Removing rows with missing values**

Note: many ML and statistical models do not tolerate NA, NaN or other missing null values.

**Remove rows with missing values / drop rows with missing values** df.dropna(axis=0, inplace=True)

**Remove rows with missing values in specific column** df.dropna(axis=0, subset=["RAIN"], inplace=True)

**Remove rows with missing values in specific columns** df.dropna(axis=0, subset=[“TEMPERATURE”, "RAIN"], inplace=True)

**Remove columns with missing values / drop columns with missing values** df.dropna(axis=1, inplace=True)

**Replacing missing values**

**Replace missing values with -1** df.fillna(value=-1, inplace=True)

You can’t specify a column using the subset parameter. To target specific columns you will need to extract them out and then apply fillna().

**NUMPY Replace values with NaN** from numpy import nandf.replace(-1, nan, inplace=True)



**SCIKIT LEARN Replace missing values with mean using SimpleImputer (imputer.fit, imputer.transform)**





There are other options for the *strategy* parameter including 'mean', 'median', 'most\_frequent', and 'constant'.

**SCIKIT LEARN Replace missing values with nearest neighbor using KNNImputer**



 





**Outliers**

We can use tools like *standard deviation (SD or* *σ [sigma])*  and *interquartile range*.

**Mean**

**Mean** mean = df.mean(axis=0)



**Standard deviation (SD, σ, sigma)**

**Standard deviation** sd = df.std(axis=0)

When calculating standard deviation with Pandas, it will be assumed to be a sample and therefore will calculate with 1 degree of freedom by default as shown in this formula:



To get a sense of how standard deviations play a role in omitting outliers, consider the graphic below. 1 standard deviation away from the mean (average) will capture 68% of the expected data points assuming a normal distribution. 2 standard deviations will capture 95%, and 3 standard deviations will capture 99.7%. With a standard deviation, The lower the standard deviation, the more aggressively outliers will be removed.



For smaller samples, cutting off at two standard deviations will be more common. This means we would declare any data on the tails outside those two standard deviations to be outliers and become candidate for removal.

Let's inspect the outliers outside two standard deviations. Multiply the standard deviation by 2 and subtract/add from the mean respectively to get the lower and upper bounds. Then we can compose a condition to identify the outliers by checking for weights less than or greater than these lower and upper bounds respectively.



**Remove outliers that fall outside the two standard deviations**

df = df[(lower < df[‘column’]) & (df['column'] < upper)]

df

**PANDAS NUMPY Remove outliers that fall outside 2.25 standard deviations:**

Don’t forget the (axis=0)**[0]** for the mean and std.



**Interquartile range outliers**

There is a lot of data that does not follow the nice bell curve shape of the normal distribution. Another way you can approach outliers in these cases is to use the Interquartile Range method, or IQR. This is the difference between the 75th and 25th percentile. When referring to the quarterly percentiles (0, 25, 50, 75, and 100). we refer to them as quartiles. A 50 percent quartile would be the middle-most value (the median), or the average of the two most-centered values.

Using the IQR, you will define a cutoff by a factor 𝑘 below or above the 25th and 75th percentile respectively. A common value for 𝑘 is 1.5, whereas a value of 3.0 would be used for more extreme cutoffs.

**Calculate percentile**

**NUMPY** In Python, we can use the **percentile() function** in NumPy to find a given percentile in a datastet.



**Get IQR:**





As you see above, the k value might be too generous for this dataset if we are looking to remove outliers. Maybe there are not extreme enough outliers in this dataset or this technique is just not warranted. But we can try to experiment lowering that k value to see how low the threshold must be before outliers removed. Below, I find a k value of 1.1 removes an outlier, with an index of 11 and weight of 54.

You can also use this technique on multidimensional data, by specifying an IQR policy for each field you want to target the removal of outliers.

**NUMPY Interquartile range outliers full walkthrough:**







[**LocalOutlierFactor**](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html)

From a machine learning perspective, you can treat outliers as a classification. If they are far away from the rest of the datapoints in a multidimensional space, they can be detected as outliers. However, this becomes less reliable on higher dimensional problems due to curse of dimensionality. By leveraging logic that measures how far neighboring data points are, we can leverage the LocalOutlierFactor.

**SCIKIT LEARN Import LocalOutlierFactor** from sklearn.neighbors import LocalOutlierFactor







**Dates and times**

Check the data type of the date columns using *df.dtypes*. It might be *object* instead of *datetime64*. If a date-column has the right data type assigned then you can extract information from that column. Like day of the week and so on.

**Change data type of the date column (parse data)**  parsed\_col = pd.to\_datetime(df[‘column’])

parsed\_col

**Extract day of week** parsed\_col.dt.dayofweek

**Format dates columns as dates while importing csv:**



**Datetime conversion**

[Dataframe pandas conventions.](https://docs.python.org/3/library/datetime.html#strftime-and-strptime-behavior)

strftime() and strptime() are used to write a datetime to a formatting string, and parse a datetime from a formatted string respectively. The format codes come from the standard C conventions. Here are a few common ones, many of which we will use in this notebook. Refer to the link above to see all format codes.



**ORDER\_DATE\_TM2 column has values like 22-Jan-22 4:08 PM. Correct them to a normal date type:**







**Get all records where day of the week is Tuesday using dt.dayofweek**



**Filter dates**



**Or filter between a specific date and time (option 1)**



And option 2



**Timezones**

**Pytz Library for timezones**

import pytz

pytz.common\_timezones

**Look up a timezone in the pytz library**

tz = pytz.timezone('Europe/Amsterdam')

tz

<DstTzInfo 'Europe/Amsterdam' LMT+0:20:00 STD>

**Check if the column has timezone information using** [**.dt.tz**](https://pandas.pydata.org/docs/reference/api/pandas.Series.dt.tz.html)



**Assign a timezone to a column**



**Convert to a different timezone**



**Convert to UTC**



**Wrangling text / data cleaning**

**Explore the topic of Regular expressions**

[**re — Regular expression operations**](https://docs.python.org/3/library/re.html)

[**An introduction to regular expressions**](https://www.oreilly.com/content/an-introduction-to-regular-expressions/)

These are the common string operations in Pandas we can use. Note that these typically accept a regular expression as a pattern, and we will cover this.

**Function Description**

count() Counts the number of instances in a pattern

contains() Returns a boolean True/False indicating whether a string contains a pattern

replace() Replaces the found patterns in a string with another specified string.

fullmatch() Determines if the entire string matches the pattern

split() Splits a string into separate strings using the pattern as the separator

extract() Finds all occurrences of a pattern and packages them into columns

findall() Finds all occurrences of a pattern and packages them into a list