hhpy

Release 0.1.0

Jan 14, 2020

Contents:

CHAPTER 1

python API reference

1.1 hhpy.main Module

1.1.1 hhpy.main.py

Contains basic calculation functions that are used in the more specialized versions of the package but can also be used on their own

1.1.2 Functions

today(date_format)	Returns today's date as string
size(byte, unit, dec)	Formats bytes as human readable string
mem_usage(pandas_obj, *args, **kwargs)	Get memory usage of a pandas object
tprint(*args, sep, **kwargs)	Wrapper for print() but with a carriage return at the end.
fprint(*args, file, sep, mode, append_sep, ...)	Write the output of print to a file instead.
elapsed_time_init()	Resets reference time for elapsed_time()
elapsed_time(do_return, ref_t)	Get the elapsed time since reference time ref_time.
total_time(i, i_max)	Estimates total time of running operation by linear extrapolation using iteration counters.
remaining_time(i, i_max)	Estimates remaining time of running operation by linear extrapolation using iteration counters.
progressbar(i, i_max, symbol, empty_symbol, ...)	Prints a progressbar for the currently running process based on iteration counters.
time_to_str(t, time_format)	Wrapper for strftime
cf_vec(x, func, *args, **kwargs)	Pandas compatible vectorize function.
round_signif_i(x, digits)	Round to significant number of digits
round_signif(x, *args, **kwargs)	Round to significant number of digits
floor_signif(x, digits)	Floor to significant number of digits
ceil_signif(x, digits)	Ceil to significant number of digits

Continued on next page

Table 1 – continued from previous page

concat_cols(df, columns, sep, to_int)	Concat a number of columns of a pandas DataFrame
list_unique(lst)	Returns unique elements from a list
list_flatten(lst)	Flatten a list of lists
list_merge(*args[, unique, flatten])	Merges n lists together
list_intersection(lst, *args)	Returns common elements of n lists
list_exclude(lst, *args)	Returns a list that includes only those elements from the first list that are not in any subsequent list.
rand(shape, lower, upper, step, seed)	A seedable wrapper for numpy.random.random_sample that allows for boundaries and steps
dict_list(*args)	Creates a dictionary of empty named lists.
append_to_dict_list(dct, append, list], inplace)	Appends to a dictionary of named lists.
is_list_like(obj)	Checks any python object to see if it is list like
force_list(*args)	Takes any python object and turns it into an iterable list.
qformat(value, int_format, float_format, . . .)	Creates a human readable representation of a generic python object
to_hdf(df, file, groupby, List[str]] = None, . . .)	saves a pandas DataFrame as h5 file, if groupby is supplied will save each group with a different key.
get_hdf_keys(file)	Reads all keys from an hdf file and returns as list
read_hdf(file, key, List[str]] = None, . . .)	read a DataFrame from hdf file

today

hhpy.main.**today**(*date_format: str = '%Y_%m_%d'*) → str

Returns today's date as string

Parameters **date_format** – The formating string for the date. Passed to strftime

Returns Formated String

Examples

```
>>> today()
'2020_01_14'
```

size

hhpy.main.**size**(*byte: int*, *unit: str = 'MB'*, *dec: int = 2*) → str

Formats bytes as human readable string

Parameters

- **byte** – The byte amount to be formated
- **unit** – The unit to display the output in, supports 'KB', 'MB', 'GB' and 'TB'
- **dec** – The number of decimals to use

Returns Formated bytes as string

Examples

```
>>> size(1024, unit='KB')
'1.0 KB'
```

```
>>> size(1024*1024*10, unit='MB')
'10.0 MB'
```

```
>>> size(10**10, unit='GB')
'9.31 GB'
```

mem_usage

hhpy.main.**mem_usage**(*pandas_obj*, **args*, ***kwargs*) → str

Get memory usage of a pandas object

Parameters

- **pandas_obj** – Pandas object to get the memory usage of
- **args** – passed to size()
- **kwargs** – passed to size()

Returns memory usage of a pandas object formated as string

Examples

```
>>> import seaborn as sns
>>> diamonds = sns.load_dataset('diamonds')
>>> mem_usage(diamonds)
'12.62 MB'
```

tprint

hhpy.main.**tprint**(**args*, *sep: str = ' '*, ***kwargs*)

Wrapper for print() but with a carriage return at the end. This results in the text being overwritten by the next print call. Can be used for progress bars and the like.

Parameters

- **args** – arguments to print
- **sep** – separator
- **kwargs** – passed to print

Returns None

Examples

```
>>> tprint('Hello World')
'Hello World'
```

```
>>> tprint(1)
>>> tprint(2)
2
```

fprint

hhpy.main.**fprint**(**args*, *file: str = '_fprint.txt'*, *sep: str = ' '*, *mode: str = 'replace'*, *append_sep: str = '\n'*, *timestamp: bool = True*, *do_print: bool = False*, *do_tprint: bool = False*)

Write the output of print to a file instead. Supports also writing to console.

Parameters

- **args** – the arguments to print
- **file** – the name of the file to print to
- **sep** – separator
- **mode** – weather to append or replace the contents of the file
- **append_sep** – if mode=='append', use this separator
- **timestamp** – weather to include a timestamp in the print statement
- **do_print** – weather to also print to console
- **do_tprint** – weather to also print to console using tprint

Returns None

Examples

The below output gets written to a file called 'fprint.txt'

```
>>> fprint('Hello World', file='fprint.txt')
```

The below output gets written both to a file and to console

```
>>> fprint('Hello World', file='fprint.txt', do_print=True)
'Hello World'
```

elapsed_time_init

hhpy.main.**elapsed_time_init**() → None

Resets reference time for elapsed_time()

Returns None

Examples

see *elapsed_time()*

elapsed_time

hhpy.main.**elapsed_time**(*do_return: bool = True*, *ref_t: datetime.datetime = None*) → datetime.timedelta

Get the elapsed time since reference time ref_time.

Parameters

- **do_return** – Whether to return or print
- **ref_t** – Reference time. If None is provided the time elapsed_time_init() was last called is used.

Returns In case of do_return: Datetime object containing the elapsed time. Else calls tprint and returns None.

Examples

```
>>> from time import sleep
>>> elapsed_time_init()
>>> sleep(1)
```

(continues on next page)

(continued from previous page)

```
>>> elapsed_time(do_return=False)
'0:00:01.0'
```

```
>>> from time import sleep
>>> elapsed_time_init()
>>> sleep(1)
>>> elapsed_time(do_return=True)
datetime.timedelta(0, 1, 1345)
```

total_time

hhpy.main.**total_time**(*i: int*, *i_max: int*) → datetime.timedelta

Estimates total time of running operation by linear extrapolation using iteration counters.

Parameters

- **i** – current iteration
- **i_max** – max iteration

Returns datetime object representing estimated total time of operation

remaining_time

hhpy.main.**remaining_time**(*i: int*, *i_max: int*) → datetime.timedelta

Estimates remaining time of running operation by linear extrapolation using iteration counters.

Parameters

- **i** – current iteration
- **i_max** – max iteration

Returns datetime object representing estimated remaining time of operation

progressbar

hhpy.main.**progressbar**(*i: int = 1*, *i_max: int = 1*, *symbol: str = '='*, *empty_symbol: str = '_'*, *mid: str = None*, *mode: str = 'perc'*, *print_prefix: str = ''*, *p_step: int = 1*, *printf: Callable = <function tprint>*, *persist: bool = False*, ***kwargs*)

Prints a progressbar for the currently running process based on iteration counters.

Parameters

- **i** – current iteration
- **i_max** – max iteration
- **symbol** – symbol that represents reached progress blocks
- **empty_symbol** – symbol that represents not yet reached progress blocks
- **mid** – what to write in the middle of the progressbar, if mid is passed mode is ignored
- **mode** – {'perc', 'total', 'elapsed'}. If perc is passed writes percentage. If 'remaining' or 'elapsed' writes remaining or elapsed time respectively. [optional]
- **print_prefix** – what to write in front of the progressbar. Useful when calling progressbar multiple times from different functions.

- **p_step** – progressbar prints one symbol (progress block) per p_step
- **printf** – Using tprint by default. Use fprint to write to file instead.
- **persist** – Whether to persist the progressbar after reaching 100 percent.
- **kwargs** – Passed to print function

Returns

time_to_str

hhpy.main.**time_to_str**(*t: datetime.datetime*, *time_format: str = '%Y-%m-%d'*) → str
Wrapper for strftime

Parameters

- **t** – datetime object
- **time_format** – time format, passed to strftime

Returns formated datetime as string

cf_vec

hhpy.main.**cf_vec**(*x: Any*, *func: Callable*, **args*, ***kwargs*) → Any
Pandas compatible vectorize function. In case a DataFrame is passed the function is applied to all columns.

Parameters

- **x** – Any vector like object
- **func** – Any function that should be vectorized
- **args** – passed to func
- **kwargs** – passed to func

Returns Vector like object

round_signif_i

hhpy.main.**round_signif_i**(*x: numpy.number*, *digits: int = 1*) → float
Round to significant number of digits

Parameters

- **x** – any number
- **digits** – integer amount of significant digits

Returns float rounded to significant digits

round_signif

hhpy.main.**round_signif**(*x: Any*, **args*, ***kwargs*) → Any
Round to significant number of digits

Parameters

- **x** – any vector like object of numbers

- **args** – passed to cf_vec
- **kwargs** – passed to cf_vec

Returns Vector like object of floats rounded to significant digits

floor_signif

hhpy.main.**floor_signif**(*x: Any*, *digits: int = 1*) → Any
Floor to significant number of digits

Parameters

- **x** – any vector like object of numbers
- **digits** – integer amount of significant digits

Returns float floored to significant digits

ceil_signif

hhpy.main.**ceil_signif**(*x: Any*, *digits: int = 1*) → Any
Ceil to significant number of digits

Parameters

- **x** – any vector like object of numbers
- **digits** – integer amount of significant digits

Returns float ceiled to significant digits

concat_cols

hhpy.main.**concat_cols**(*df: pandas.core.frame.DataFrame*, *columns: list*, *sep: str = '_'*, *to_int: bool = False*) → pandas.core.series.Series
Concat a number of columns of a pandas DataFrame

Parameters

- **df** – Pandas DataFrame
- **columns** – Names of the columns to be concat
- **sep** – Separator
- **to_int** – If true: Converts columns to int before concatting

Returns Pandas Series containing the concat columns

list_unique

hhpy.main.**list_unique**(*lst: list*) → list
Returns unique elements from a list

Parameters **lst** – any list like object

Returns a list

list_flatten

hhpy.main.**list_flatten**(*lst: list*) → list
Flatten a list of lists

Parameters **lst** – list of lists

Returns flattened list

list_merge

hhpy.main.**list_merge**(**args*, *unique=True*, *flatten=False*) → list
Merges n lists together

Parameters

- **args** – The lists to be merged together
- **unique** – if True then duplicate elements will be dropped
- **flatten** – if True then the individual lists will be flatten before merging

Returns The merged list

list_intersection

hhpy.main.**list_intersection**(*lst: list*, **args*) → list
Returns common elements of n lists

Parameters

- **lst** – the first list
- **args** – the subsequent lists

Returns the list of common elements

list_exclude

hhpy.main.**list_exclude**(*lst: list*, **args*) → list
Returns a list that includes only those elements from the first list that are not in any subsequent list. Can also be called with non list args, then those elements are removed.

Parameters

- **lst** – the list to exclude from
- **args** – the subsequent lists

Returns the filtered list

rand

hhpy.main.**rand**(*shape: tuple = None*, *lower: int = None*, *upper: int = None*, *step: int = None*, *seed: int = None*) → numpy.array
A seedable wrapper for numpy.random.random_sample that allows for boundaries and steps

Parameters

- **shape** – A tuple containing the shape of the desired output array

- **lower** – Lower bound of random numbers
- **upper** – Upper bound of random numbers
- **step** – Minimum step between random numbers
- **seed** – Random Seed

Returns Numpy Array

dict_list

hhpy.main.**dict_list**(**args*) → dict

Creates a dictionary of empty named lists. Useful for iteratively creating a pandas DataFrame

Parameters **args** – The names of the lists

Returns Dictionary of empty named lists

append_to_dict_list

hhpy.main.**append_to_dict_list**(*dct: dict, append: Union[dict, list], inplace: bool = True*) → Optional[dict]

Appends to a dictionary of named lists. Useful for iteratively creating a pandas DataFrame.

Parameters

- **dct** – Dictionary to append to
- **append** – List or dictionary of values to append
- **inplace** – Modify inplace or return modified copy

Returns None if inplace, else modified dictionary

is_list_like

hhpy.main.**is_list_like**(*obj: Any*) → bool

Checks any python object to see if it is list like

Parameters **obj** – Any python object

Returns Boolean

force_list

hhpy.main.**force_list**(**args*) → list

Takes any python object and turns it into an iterable list.

Parameters **args** – Any python object

Returns List

qformat

hhpy.main.**qformat**(*value: Any*, *int_format: str = ''*, *float_format: str = '.2f'*, *datetime_format: str = '%Y-%m-%d'*, *sep: str = ' - '*, *key_sep: str = ': '*, *print_key: bool = True*) → str

Creates a human readable representation of a generic python object

Parameters

- **value** – Any python object
- **int_format** – Format string for integer
- **float_format** – Format string for float
- **datetime_format** – Format string for datetime
- **sep** – Separator
- **key_sep** – Separator used between key and value if print_key is True
- **print_key** – Whether to print keys as well as values (if object has keys)

Returns Formated string

to_hdf

hhpy.main.**to_hdf**(*df: pandas.core.frame.DataFrame*, *file: str*, *groupby: Union[str, List[str]] = None*, *key: str = None*, *replace: bool = False*, *do_print=True*, ***kwargs*) → None

saves a pandas DataFrame as h5 file, if groupby is supplied will save each group with a different key. Needs with groupby OR key to be supplied. Extends on pandas.DataFrame.to_hdf.

Parameters

- **df** – DataFrame to save
- **file** – filename to save the DataFrame as
- **groupby** – if supplied will save each sub-DataFrame as a different key. [optional]
- **key** – The key to write as. Ignored if groupby is supplies.
- **replace** – Whether to replace or append to existing files. Defaults to append. [optional]
- **do_print** – Whether to print intermediate steps to console [optional]
- **kwargs** – Other keyword arguments passed to pd.DataFrame.to_hdf [optional]

Returns None

get_hdf_keys

hhpy.main.**get_hdf_keys**(*file: str*) → List[str]

Reads all keys from an hdf file and returns as list

Parameters **file** – The path of the file to read the keys of

Returns List of keys

read_hdf

hhpy.main.**read_hdf**(*file: str*, *key: Union[str, List[str]] = None*, *sample: int = None*, *random_state: int = None*, *do_print: bool = True*, *catch_error: bool = True*) → pandas.core.frame.DataFrame

read a DataFrame from hdf file

Parameters

- **file** – The path to the file to read from
- **key** – The key(s) to read, if not specified all keys are read [optional]
- **sample** – If specified will read sample keys at random from the file, ignored if key is specified [optional]
- **random_state** – Random state for sample [optional]
- **do_print** – Whether to print intermediate steps [optional]
- **catch_error** – Whether to catch errors when reading [optional]

Returns pandas DataFrame

1.2 hhpy.ds Module

1.2.1 hhpy.ds.py

Contains DataScience functions extending on pandas and sklearn

1.2.2 Functions

optimize_pd(df, c_int, c_float, c_cat, cat_frac)	optimize memory usage of a pandas df, automatically downcast all var types and converts objects to categories
get_df_corr(df, target, groupby, list] = None)	returns a pandas DataFrame containing all pearson correlations in a melted format
drop_zero_cols(df)	Drop columns with all 0 or None Values from DataFrame.
get_duplicate_indices(df)	Returns duplicate indices from a pandas DataFrame
get_duplicate_cols(df)	Returns names of duplicate columns from a pandas DataFrame
drop_duplicate_indices(df)	Drop duplicate indices from pandas DataFrame
drop_duplicate_cols(df)	Drop duplicate columns from pandas DataFrame
change_span(s, steps)	return a True/False series around a changepoint, used for filtering stepwise data series in a pandas df must be properly sorted!
outlier_to_nan(df, col, groupby, . . .)	this algorithm cuts off all points whose DELTA (avg diff to the prev and next point) is outside of the n std range
butter_pass_filter(data, cutoff, fs, order, . . .)	Implementation of a highpass / lowpass filter using scipy.signal.butter
pass_by_group(df, col, groupby, list], . . .)	allows applying a butter_pass filter by group
lfit(x, str], y, str] = None, w, . . .)	quick linear fit with numpy
qf(df, fltr, pandas.core.series.Series, . . .)	quickly filter a DataFrame based on equal criteria.

Continued on next page

Table 2 – continued from previous page

quantile_split(s, n, signif, na_to_med)	splits a numerical column into n quantiles.
acc(y_true, str], y_pred, str], df)	calculate accuracy for a categorical label
rel_acc(y_true, str], y_pred, str], df, ...)	relative accuracy of the prediction in comparison to predicting everything as the most common group :param y_true: true values as name of df or vector data :param y_pred: predicted values as name of df or vector data :param df: pandas DataFrame containing true and predicted values [optional] :param target_class: name of the target class, by default the most common one is used [optional] :return: accuracy difference as percent
cm(y_true, str], y_pred, str], df)	confusion matrix from pandas df :param y_true: true values as name of df or vector data :param y_pred: predicted values as name of df or vector data :param df: pandas DataFrame containing true and predicted values [optional] :return: Confusion matrix as pandas DataFrame
f1_pr(y_true, str], y_pred, str], df, ...)	get f1 score, true positive, true negative, missed positive and missed negative rate
f_score(y_true, str], y_pred, str], df, ...)	generic scoring function base on pandas DataFrame.
r2(*args, **kwargs)	wrapper for f_score using sklearn.metrics.r2_score
rmse(*args, **kwargs)	wrapper for f_score using numpy.sqrt(skearn.metrics.mean_squared_error)
mae(*args, **kwargs)	wrapper for f_score using skearn.metrics.mean_absolute_error
stdae(*args, **kwargs)	wrapper for f_score using the standard deviation of the absolute error
medae(*args, **kwargs)	wrapper for f_score using skearn.metrics.median_absolute_error
corr(*args, **kwargs)	wrapper for f_score using pandas.Series.corr
df_score(df, y_true, str], pred_suffix, ...)	creates a DataFrame displaying various kind of scores
rmsd(x, df, group, return_df_paired, ...)	calculated the weighted root mean squared difference for a reference columns x by a specific group
df_rmsd(x, df, groups, str] = None, hue, ...)	calculate rmsd for reference column x with multiple other columns and return as DataFrame
df_p(x, group, df, hue, agg_func, agg, ...)	returns a DataFrame with the p value.
df_split(df, split_by, str], return_type, ...)	Split a pandas DataFrame by column value and returns a list or dict
mahalanobis(point, ...)	Calculates the Mahalanobis distance for a single point or a DataFrame of points
top_n(s, n, w)	select n elements form a categorical pandas series with the highest counts
top_n_coding(s, n, other_name, na_to_other, w)	returns a modified version of the pandas series where all elements not in top_n become recoded as 'other'
k_split(df, k, groupby, str] = None, sortby, ...)	splits a DataFrame into k (equal sized) parts that can be used for train test splitting or k_cross splitting

optimize_pd

hhpy.ds.**optimize_pd**(*df: pandas.core.frame.DataFrame*, *c_int: bool = True*, *c_float: bool = True*, *c_cat: bool = True*, *cat_frac: bool = 0.5*) → pandas.core.frame.DataFrame

optimize memory usage of a pandas df, automatically downcast all var types and converts objects to categories

Parameters

- **df** – pandas DataFrame to be optimized. Other objects are implicitly cast to DataFrame
- **c_int** – whether to downcast integers
- **c_float** – whether to downcast floats
- **c_cat** – whether to cast objects to categories. Uses cat_frac as condition
- **cat_frac** – if c_cat is True and the column has less than cat_frac unique values it will be cast to category

Returns the optimized pandas DataFrame

get_df_corr

hhpy.ds.**get_df_corr**(*df: pandas.core.frame.DataFrame*, *target: str = None*, *groupby: Union[str, list] = None*) → pandas.core.frame.DataFrame

returns a pandas DataFrame containing all pearson correlations in a melted format

Parameters

- **df** – input pandas DataFrame. Other objects are implicitly cast to DataFrame
- **target** – if target is specified: returns only correlations that involve the target column
- **groupby** – if groupby is specified: returns correlations for each level of the group

Returns pandas DataFrame containing all pearson correlations in a melted format

drop_zero_cols

hhpy.ds.**drop_zero_cols**(*df: pandas.core.frame.DataFrame*) → pandas.core.frame.DataFrame

Drop columns with all 0 or None Values from DataFrame. Useful after applying one hot encoding.

Parameters **df** – pandas DataFrame

Returns pandas DataFrame without 0 columns.

get_duplicate_indices

hhpy.ds.**get_duplicate_indices**(*df: pandas.core.frame.DataFrame*) → Sequence[T_co]

Returns duplicate indices from a pandas DataFrame

Parameters **df** – pandas DataFrame

Returns List of indices that are duplicate

get_duplicate_cols

hhpy.ds.**get_duplicate_cols**(*df: pandas.core.frame.DataFrame*) → Sequence[T_co]

Returns names of duplicate columns from a pandas DataFrame

Parameters **df** – pandas DataFrame

Returns List of column names that are duplicate

drop_duplicate_indices

hhpy.ds.**drop_duplicate_indices**(*df: pandas.core.frame.DataFrame*) → pandas.core.frame.DataFrame

Drop duplicate indices from pandas DataFrame

Parameters **df** – pandas DataFrame

Returns pandas DataFrame without the duplicates indices

drop_duplicate_cols

hhpy.ds.**drop_duplicate_cols**(*df: pandas.core.frame.DataFrame*) → pandas.core.frame.DataFrame

Drop duplicate columns from pandas DataFrame

Parameters **df** – pandas DataFrame

Returns pandas DataFrame without the duplicates columns

change_span

hhpy.ds.**change_span**(*s: pandas.core.series.Series*, *steps: int = 5*) → pandas.core.series.Series

return a True/False series around a changepoint, used for filtering stepwise data series in a pandas df must be properly sorted!

Parameters

- **s** – pandas Series or similar
- **steps** – number of steps around the changepoint to flag as true

Returns pandas Series of dtype Boolean

outlier_to_nan

hhpy.ds.**outlier_to_nan**(*df: pandas.core.frame.DataFrame*, *col: str*, *groupby: Union[list, str] = None*, *std_cutoff: numpy.number = 3*, *reps: int = 1*, *do_print: bool = False*) → pandas.core.frame.DataFrame

this algorithm cuts off all points whose DELTA (avg diff to the prev and next point) is outside of the n std range

Parameters

- **df** – pandas DataFrame
- **col** – column to be filtered
- **groupby** – if provided: applies std filter by group
- **std_cutoff** – the number of standard deviations outside of which to set values to None
- **reps** – how many times to repeat the algorithm
- **do_print** – whether to print steps to console

Returns pandas DataFrame with outliers set to nan

butter_pass_filter

hhpy.ds.**butter_pass_filter**(*data: pandas.core.series.Series*, *cutoff: int*, *fs: int*, *order: int*, *btype: str = None*, *shift: bool = False*)

Implementation of a highpass / lowpass filter using scipy.signal.butter

Parameters

- **data** – pandas Series or 1d numpy Array
- **cutoff** – cutoff
- **fs** – critical frequencies
- **order** – order of the fit
- **btype** – The type of filter. Passed to scipy.signal.butter. Default is 'lowpass'. One of {'lowpass', 'highpass', 'bandpass', 'bandstop'}
- **shift** – whether to shift the data to start at 0

Returns 1d numpy array containing the filtered data

pass_by_group

hhpy.ds.**pass_by_group**(*df: pandas.core.frame.DataFrame*, *col: str*, *groupby: Union[str, list]*, *btype: str*, *shift: bool = False*, *cutoff: int = 1*, *fs: int = 20*, *order: int = 5*)

allows applying a butter_pass filter by group

Parameters

- **df** – pandas DataFrame
- **col** – column to filter
- **groupby** – columns to groupby
- **btype** – The type of filter. Passed to scipy.signal.butter. Default is 'lowpass'. One of {'lowpass', 'highpass', 'bandpass', 'bandstop'}
- **shift** – shift: whether to shift the data to start at 0
- **cutoff** – cutoff
- **fs** – critical frequencies
- **order** – order of the filter

Returns filtered DataFrame

lfit

hhpy.ds.**lfit**(*x: Union[pandas.core.series.Series, str]*, *y: Union[pandas.core.series.Series, str] = None*, *w: Union[pandas.core.series.Series, str] = None*, *df: pandas.core.frame.DataFrame = None*, *groupby: Union[list, str] = None*, *do_print: bool = True*, *catch_error: bool = False*, *return_df: bool = False*, *extrapolate: bool = None*)

quick linear fit with numpy

Parameters

- **x** – names of x variables in df or vector data, if y is None treated as target and fit against the index
- **y** – names of y variables in df or vector data [optional]

- **w** – names of weight variables in df or vector data [optional]
- **df** – pandas DataFrame containing x,y,w data [optional]
- **groupby** – If specified the linear fit is applied by group [optional]
- **do_print** – whether to print steps to console
- **catch_error** – whether to keep going in case of error [optional]
- **return_df** – whether to return a DataFrame or Series [optional]
- **extrapolate** – how many iteration to extrapolate [optional]

Returns if return_df is True: pandas DataFrame, else: pandas Series

qf

hhpy.ds.**qf**(*df: pandas.core.frame.DataFrame, fltr: Union[pandas.core.frame.DataFrame, pandas.core.series.Series, Mapping[KT, VT_co]], remove_unused_categories: bool = True, reset_index: bool = False*)

quickly filter a DataFrame based on equal criteria. All columns of fltr present in df are filtered to be equal to the first entry in filter_df.

Parameters

- **df** – pandas DataFrame to be filtered
- **fltr** – filter condition as DataFrame or Mapping or Series
- **remove_unused_categories** – whether to remove unused categories from categorical dtype after filtering
- **reset_index** – whether to reset index after filtering

Returns filtered pandas DataFrame

quantile_split

hhpy.ds.**quantile_split**(*s: pandas.core.series.Series, n: int, signif: int = 2, na_to_med: bool = False*)

splits a numerical column into n quantiles. Useful for mapping numerical columns to categorical columns

Parameters

- **s** – pandas Series to be split
- **n** – number of quantiles to split into
- **signif** – number of significant digits to round to
- **na_to_med** – whether to fill na values with median values

Returns pandas Series of dtype category

acc

hhpy.ds.**acc**(*y_true: Union[pandas.core.series.Series, str], y_pred: Union[pandas.core.series.Series, str], df: pandas.core.frame.DataFrame = None*) → float

calculate accuracy for a categorical label

Parameters

- **y_true** – true values as name of df or vector data
- **y_pred** – predicted values as name of df or vector data
- **df** – pandas DataFrame containing true and predicted values [optional]

Returns accuracy a percentage

rel_acc

hhpy.ds.**rel_acc**(*y_true: Union[pandas.core.series.Series, str], y_pred: Union[pandas.core.series.Series, str], df: pandas.core.frame.DataFrame = None, target_class: str = None*)

relative accuracy of the prediction in comparison to predicting everything as the most common group :param y_true: true values as name of df or vector data :param y_pred: predicted values as name of df or vector data :param df: pandas DataFrame containing true and predicted values [optional] :param target_class: name of the target class, by default the most common one is used [optional] :return: accuracy difference as percent

cm

hhpy.ds.**cm**(*y_true: Union[pandas.core.series.Series, str], y_pred: Union[pandas.core.series.Series, str], df: pandas.core.frame.DataFrame = None*) → pandas.core.frame.DataFrame

confusion matrix from pandas df :param y_true: true values as name of df or vector data :param y_pred: predicted values as name of df or vector data :param df: pandas DataFrame containing true and predicted values [optional] :return: Confusion matrix as pandas DataFrame

f1_pr

hhpy.ds.**f1_pr**(*y_true: Union[pandas.core.series.Series, str], y_pred: Union[pandas.core.series.Series, str], df: pandas.core.frame.DataFrame = None, target: str = None, factor: int = 100*) → pandas.core.frame.DataFrame

get f1 score, true positive, true negative, missed positive and missed negative rate

Parameters

- **y_true** – true values as name of df or vector data
- **y_pred** – predicted values as name of df or vector data
- **df** – pandas DataFrame containing true and predicted values [optional]
- **target** – level for which to return the rates, by default all levels are returned [optional]
- **factor** – factor by which to scale results, default 100 [optional]

Returns pandas DataFrame containing f1 score, true positive, true negative, missed positive and missed negative rate

f_score

hhpy.ds.**f_score**(*y_true: Union[pandas.core.series.Series, str], y_pred: Union[pandas.core.series.Series, str], df: pandas.core.frame.DataFrame = None, dropna: bool = False, f: Callable = <function r2_score>, groupby: Union[list, str] = None, f_name: str = None*) → Union[pandas.core.frame.DataFrame, float]

generic scoring function base on pandas DataFrame.

Parameters

- **y_true** – true values as name of df or vector data
- **y_pred** – predicted values as name of df or vector data
- **df** – pandas DataFrame containing true and predicted values [optional]
- **dropna** – whether to dropna values [optional]
- **f** – scoreing function to apply, default is sklearn.metrics.r2_score, should return a scalar value. [optional]
- **groupby** – if supplied then the result is returned for each group level [optional]
- **f_name** – name of the scoreing function, by default uses .__name__ property of fuction [optional]

Returns if groupby is supplied: pandas DataFrame, else: scalar value

r2

hhpy.ds.**r2**(**args*, ***kwargs*) → Union[pandas.core.frame.DataFrame, float]
wrapper for f_score using sklearn.metrics.r2_score

Parameters

- **args** – passed to f_score
- **kwargs** – passed to f_score

Returns if groupby is supplied: pandas DataFrame, else: scalar value

rmse

hhpy.ds.**rmse**(**args*, ***kwargs*) → Union[pandas.core.frame.DataFrame, float]
wrapper for f_score using numpy.sqrt(skearn.metrics.mean_squared_error)

Parameters

- **args** – passed to f_score
- **kwargs** – passed to f_score

Returns if groupby is supplied: pandas DataFrame, else: scalar value

mae

hhpy.ds.**mae**(**args*, ***kwargs*) → Union[pandas.core.frame.DataFrame, float]
wrapper for f_score using skearn.metrics.mean_absolute_error

Parameters

- **args** – passed to f_score
- **kwargs** – passed to f_score

Returns if groupby is supplied: pandas DataFrame, else: scalar value

stdae

hhpy.ds.**stdae**(**args*, ***kwargs*) → Union[pandas.core.frame.DataFrame, float]
wrapper for f_score using the standard deviation of the absolute error

Parameters

- **args** – passed to f_score
- **kwargs** – passed to f_score

Returns if groupby is supplied: pandas DataFrame, else: scalar value

medae

hhpy.ds.**medae**(**args*, ***kwargs*) → Union[pandas.core.frame.DataFrame, float]
wrapper for f_score using skearn.metrics.median_absolute_error

Parameters

- **args** – passed to f_score
- **kwargs** – passed to f_score

Returns if groupby is supplied: pandas DataFrame, else: scalar value

corr

hhpy.ds.**corr**(**args*, ***kwargs*) → Union[pandas.core.frame.DataFrame, float]
wrapper for f_score using pandas.Series.corr

Parameters

- **args** – passed to f_score
- **kwargs** – passed to f_score

Returns if groupby is supplied: pandas DataFrame, else: scalar value

df_score

hhpy.ds.**df_score**(*df: pandas.core.frame.DataFrame, y_true: Union[List[str], str], pred_suffix: list = None, scores: List[Callable] = None, pivot: bool = True, scale: Union[dict, list, int] = None, groupby: Union[list, str] = None*) → pandas.core.frame.DataFrame
creates a DataFrame displaying various kind of scores

Parameters

- **df** – pandas DataFrame containing the true, pred data
- **y_true** – name of the true variable inside df
- **pred_suffix** – name of the predicted variable suffixes. Supports multiple predictions. By default assumed suffix 'pred' [optional]
- **scores** – scoring functions to be used [optional]
- **pivot** – whether to pivot the DataFrame for easier readability [optional]
- **scale** – a scale for multiplying the scores, default 1 [optional]
- **groupby** – if supplied then the scores are calculated by group [optional]

Returns pandas DataFrame containing al the scores

rmsd

hhpy.ds.**rmsd**(*x: str*, *df: pandas.core.frame.DataFrame*, *group: str*, *return_df_paired: bool = False*, *agg_func: str = 'median'*, *standardize: bool = False*, *to_abs: bool = False*) → Union[float, pandas.core.frame.DataFrame]

calculated the weighted root mean squared difference for a reference columns x by a specific group

Parameters

- **x** – name of the column to calculate the rmsd for
- **df** – pandas DataFrame
- **group** – groups for which to calculate the rmsd
- **return_df_paired** – whether to return the paired DataFrame
- **agg_func** – which aggregation to use for the group value, passed to pd.DataFrame.agg
- **standardize** – whether to apply Standardization before calculating the rmsd
- **to_abs** – whether to cast x to abs before calculating the rmsd

Returns if return_df_paired pandas DataFrame, else rmsd as float

df_rmsd

hhpy.ds.**df_rmsd**(*x: str*, *df: pandas.core.frame.DataFrame*, *groups: Union[list, str] = None*, *hue: str = None*, *hue_order: list = None*, *sort_by_hue: bool = True*, *n_quantiles: int = 10*, *include_rmsd: bool = True*, ***kwargs*)

calculate rmsd for reference column x with multiple other columns and return as DataFrame

Parameters

- **x** – name of the column to calculate the rmsd for
- **df** – pandas DataFrame containing the data
- **groups** – groups to calculate the rmsd or, defaults to all other columns in the DataFrame [optional]
- **hue** – further calculate the rmsd for each hue level [optional]
- **hue_order** – sort the hue levels in this order [optional]
- **sort_by_hue** – sort the values by hue rather than by group [optional]
- **n_quantiles** – numeric columns will be automatically split into this many quantiles [optional]
- **include_rmsd** – if False provide only a grouped DataFrame but don't actually calculate the rmsd, you can use include_rmsd=False to save computation time if you only need the maxperc (used in plotting)
- **kwargs** – passed to rmsd

Returns None

df_p

hhpy.ds.**df_p**(*x: str*, *group: str*, *df: pandas.core.frame.DataFrame*, *hue: str = None*, *agg_func: str = 'mean'*, *agg: bool = False*, *n_quantiles: int = 10*)

returns a DataFrame with the p value. See hypothesis testing. :param x: name of column to evaluate :param group: name of grouping column :param df: pandas DataFrame :param hue: further split by hue level :param agg_func: standard agg function, passed to pd.DataFrame.agg :param agg: whether to include standard aggregation :param n_quantiles: numeric columns will be automatically split into this many quantiles [optional] :return: pandas DataFrame containing p values

df_split

hhpy.ds.**df_split**(*df: pandas.core.frame.DataFrame, split_by: Union[List[str], str], return_type: str = 'dict', print_key: bool = False, sep: str = '_', key_sep: str = '=='*) → Union[list, dict]

Split a pandas DataFrame by column value and returns a list or dict

Parameters

- **df** – pandas DataFrame to be split
- **split_by** – Column(s) to split by, creates a sub-DataFrame for each level
- **return_type** – one of ['list', 'dict'], if list returns a list of sub-DataFrame, if dict returns a dictionary with each level as keys
- **print_key** – whether to include the column names in the key labels
- **sep** – separator to use in the key labels between columns
- **key_sep** – separator to use in the key labels between key and value

Returns see return_type

mahalanobis

hhpy.ds.**mahalanobis**(*point: Union[pandas.core.frame.DataFrame, pandas.core.series.Series, numpy.ndarray], df: pandas.core.frame.DataFrame = None, params: List[str] = None, do_print: bool = True*) → Union[float, List[float]]

Calculates the Mahalanobis distance for a single point or a DataFrame of points

Parameters

- **point** – The point(s) to calculate the Mahalanobis distance for
- **df** – The reference DataFrame against which to calculate the Mahalanobis distance
- **params** – The columns to calculate the Mahalanobis distance for
- **do_print** – Whether to print intermediate steps to the console

Returns if a single point is passed: Mahalanobis distance as float, else a list of floats

top_n

hhpy.ds.**top_n**(*s: Sequence[T_co], n: int, w: Optional[Sequence[T_co]] = None*) → list

select n elements form a categorical pandas series with the highest counts

Parameters

- **s** – pandas Series to select from
- **n** – how many elements to return
- **w** – weights, if given the weights are summed instead of just counting entries in s [optional]

Returns List of top n elements

top_n_coding

hhpy.ds.**top_n_coding**(*s: Sequence[T_co], n: int, other_name: str = 'other', na_to_other: bool = False, w: Optional[Sequence[T_co]] = None*) → pandas.core.series.Series

returns a modified version of the pandas series where all elements not in top_n become recoded as 'other'

Parameters

- **s** – pandas Series to adjust
- **n** – how many elements to keep
- **other_name** – name of the other element [optional]
- **na_to_other** – whether to cast missing elements to other [optional]
- **w** – weights, if given the weights are summed instead of just counting entries in s [optional]

Returns adjusted pandas Series

k_split

hhpy.ds.**k_split**(*df: pandas.core.frame.DataFrame, k: int = 5, groupby: Union[Sequence[T_co], str] = None, sortby: Union[Sequence[T_co], str] = None, random_state: int = None, do_print: bool = True, return_type: Union[str, int] = 1*) → Union[pandas.core.series.Series, tuple]

splits a DataFrame into k (equal sized) parts that can be used for train test splitting or k_cross splitting

Parameters

- **df** – pandas DataFrame to be split
- **k** – how many (equal sized) parts to split the DataFrame into [optional]
- **groupby** – passed to pandas.DataFrame.groupby before splitting, ensures that each group will be represented equally in each split part [optional]
- **sortby** – if True the DataFrame is ordered by these column(s) and then sliced into parts from the top if False the DataFrame is sorted randomly before slicing [optional]
- **random_state** – random_state to be used in random sorting, ignore if sortby is True [optional]
- **do_print** – whether to print steps to console [optional]
- **return_type** – if one of ['Series', 's'] returns a pandas Series containing the k indices range(k) if a positive integer < k returns tuple of shape (df_train, df_test) where the return_type'th part is equal to df_test and the other parts are equal to df_train

Returns depending on return_type either a pandas Series or a tuple

1.3 hhpy.ipython Module

1.3.1 hhpy.ipython.py

Contains convenience wrappers for ipython

1.3.2 Functions

wide_notebook(width)	makes the jupyter notebook wider by appending html code to change the width,
hide_code()	hides the code and introduces a toggle button
display_full(*args, **kwargs)	wrapper to display a pandas DataFrame with all rows and columns
pd_display(*args[, number_format, full])	wrapper to display a pandas DataFrame with a specified number format
display_df(df[, int_format, float_format, ...])	Wrapper to display a pandas DataFrame with separate options for int / float, also adds an option to exclude columns
highlight_max(df, color)	highlights the largest value in each column of a pandas DataFrame
highlight_min(df, color)	highlights the smallest value in each column of a pandas DataFrame
highlight_max_min(df, max_color, min_color)	highlights the largest and smallest value in each column of a pandas DataFrame

wide_notebook

hhpy.ipython.**wide_notebook**(*width: int = 90*)

makes the jupyter notebook wider by appending html code to change the width, based on https://stackoverflow.com/questions/21971449/how-do-i-increase-the-cell-width-of-the-jupyter-ipython-notebook-in-my-browser

Param width in percent, default 90 [optional]

Returns None

hide_code

hhpy.ipython.**hide_code**()

hides the code and introduces a toggle button based on https://stackoverflow.com/questions/27934885/how-to-hide-code-from-cells-in-ipython-notebook-visualized -with-nbviewer

Returns None

display_full

hhpy.ipython.**display_full**(**args*, ***kwargs*)

wrapper to display a pandas DataFrame with all rows and columns

Parameters

- **args** – passed to display
- **kwargs** – passed to display

Returns None

pd_display

hhpy.ipython.**pd_display**(**args*, *number_format='{:,.2f}'*, *full=True*, ***kwargs*)

wrapper to display a pandas DataFrame with a specified number format

Parameters

- **args** – passed to display
- **number_format** – the number format to apply
- **full** – whether to show all rows and columns or keep default behaviour
- **kwargs** – passed to display

Returns None

display_df

hhpy.ipython.**display_df**(*df*, *int_format=','*, *float_format='.2f'*, *exclude=None*, *full=True*, ***kwargs*)

Wrapper to display a pandas DataFrame with separate options for int / float, also adds an option to exclude columns

Parameters

- **df** – pandas DataFrame to display
- **int_format** – format for integer columns
- **float_format** – format for float columns
- **exclude** – columns to exclude
- **full** – whether to show all rows and columns or keep default behaviour
- **kwargs** – passed to display

Returns None

highlight_max

hhpy.ipython.**highlight_max**(*df: pandas.core.frame.DataFrame*, *color: str = 'xkcd:cyan'*) → pandas.core.frame.DataFrame

highlights the largest value in each column of a pandas DataFrame

Parameters

- **df** – pandas DataFrame
- **color** – color used for highlighting

Returns the pandas DataFrame with the style applied to it

highlight_min

hhpy.ipython.**highlight_min**(*df: pandas.core.frame.DataFrame*, *color: str = 'xkcd:light red'*) → pandas.core.frame.DataFrame

highlights the smallest value in each column of a pandas DataFrame

Parameters

- **df** – pandas DataFrame
- **color** – color used for highlighting

Returns the pandas DataFrame with the style applied to it

highlight_max_min

hhpy.ipython.**highlight_max_min**(*df: pandas.core.frame.DataFrame*, *max_color: str = 'xkcd:cyan'*, *min_color: str = 'xkcd:light red'*)

highlights the largest and smallest value in each column of a pandas DataFrame

Parameters

- **df** – pandas DataFrame
- **max_color** – color used for highlighting largest value
- **min_color** – color used for highlighting smallest value

Returns the pandas DataFrame with the style applied to it

1.4 hhpy.modelling Module

1.4.1 hhpy.modelling.py

Contains a model class that is based on pandas DataFrames and wraps around sklearn and other frameworks to provide convenient train test functions.

1.4.2 Functions

dict_to_model(dic, VT_co])	restore a Model object from a dictionary
force_model(model, Mapping[KT, VT_co]])	takes any Model, model object or dictionary and converts to Model
get_coefs(model, y, str])	get coefficients of a linear regression in a sorted data frame
get_feature_importance(model, predictors, ...)	get feature importance of a decision tree like model in a sorted data frame

dict_to_model

hhpy.modelling.**dict_to_model**(*dic: Mapping[KT, VT_co]*) → hhpy.modelling.Model

restore a Model object from a dictionary

Parameters **dic** – dictionary containing the model definition

Returns Model

force_model

hhpy.modelling.**force_model**(*model: Union[object, Mapping[KT, VT_co]]*) → hhpy.modelling.Model

takes any Model, model object or dictionary and converts to Model

Parameters **model** – Mapping or object containing a model

Returns Model

get_coefs

hhpy.modelling.**get_coefs**(*model: object, y: Union[Sequence[T_co], str]*)

get coefficients of a linear regression in a sorted data frame

Parameters

- **model** – model object
- **y** – name of the coefficients

Returns pandas DataFrame containing the coefficient names and values

get_feature_importance

hhpy.modelling.**get_feature_importance**(*model: object, predictors: Union[Sequence[T_co], str], features_to_sum: Mapping[KT, VT_co] = None*) → pandas.core.frame.DataFrame

get feature importance of a decision tree like model in a sorted data frame

Parameters

- **model** – model object
- **predictors** – names of the predictors properly sorted
- **features_to_sum** – if you want to sum features please provide name mappings

Returns pandas DataFrame containing the feature importances

1.4.3 Classes

Model(model, name, X_ref, str] = None, ...)	A unified modeling class that is extended from sklearn, accepts any model that implements .fit and .predict
Models(*args, name, df, X_ref, str] = None, ...)	Collection of Models that allow for fitting and predicting with multiple Models at once, comparing accuracy and creating Ensembles

Model

class hhpy.modelling.**Model**(*model: object = None, name: str = 'pred', X_ref: Union[Sequence[T_co], str] = None, y_ref: Union[Sequence[T_co], str] = None*)

Bases: hhpy.modelling._BaseModel

A unified modeling class that is extended from sklearn, accepts any model that implements .fit and .predict

Parameters

- **model** – Any model object that implements .fit and .predict
- **name** – Name of the model, used for naming columns [optional]
- **X_ref** – List of features (predictors) used for training the model
- **y_ref** – List of labels (targets) to be predicted

Methods Summary

fit(X, Sequence[T_co], str] = None, y, ...)	generalized fit method extending on model.fit
predict([X, df, return_type])	Generalized predict method based on model.predict

Methods Documentation

fit(*X: Union[numpy.ndarray, Sequence[T_co], str] = None, y: Union[numpy.ndarray, Sequence[T_co], str] = None, df: pandas.core.frame.DataFrame = None, dropna: bool = True, X_test: Union[numpy.ndarray, Sequence[T_co], str] = None, y_test: Union[numpy.ndarray, Sequence[T_co], str] = None, df_test: pandas.core.frame.DataFrame = None*) → None
generalized fit method extending on model.fit

Parameters

- **X** – The feature (predictor) data used for training as DataFrame, np.array or column names
- **y** – The label (target) data used for training as DataFrame, np.array or column names
- **df** – Pandas DataFrame containing the training data, optional if array like data is passed for X/y
- **dropna** – Whether to drop rows containing NA in the training data [optional]
- **X_test** – The feature (predictor) data used for testing as DataFrame, np.array or column names
- **y_test** – The label (target) data used for testing as DataFrame, np.array or column names
- **df_test** – Pandas DataFrame containing the testing data, optional if array like data is passed for X/y test

Returns None

predict(*X=None, df=None, return_type='y'*) → Union[pandas.core.series.Series, pandas.core.frame.DataFrame]
Generalized predict method based on model.predict

Parameters

- **X** – The feature (predictor) data used for training as DataFrame, np.array or column names
- **df** – Pandas DataFrame containing the training and testing data. Can be saved to the Model object or supplied on an as needed basis.
- **return_type** – one of ['y', 'df', 'DataFrame'], if 'y' returns a pandas Series / DataFrame with only the predictions, if one of 'df','DataFrame' returns the full DataFrame with predictions added

Returns see return_type

Models

class hhpy.modelling.**Models**(**args, name: str = None, df: pandas.core.frame.DataFrame = None, X_ref: Union[Sequence[T_co], str] = None, y_ref: Union[Sequence[T_co], str] = None, scaler_X: object = None, scaler_y: object = None, printf: Callable = <function tprint>*)

Bases: hhpy.modelling._BaseModel

Collection of Models that allow for fitting and predicting with multiple Models at once, comparing accuracy and creating Ensembles

Parameters

- **args** – multiple Model objects that will form a Models Collection
- **name** – name of the collection
- **df** – Pandas DataFrame containing the training and testing data. Can be saved to the Model object or supplied on an as needed basis.
- **X_ref** – List of features (predictors) used for training the model
- **y_ref** – List of labels (targets) to be predicted
- **scaler_X** – Scalar object that implements .transform and .inverse_transform, applied to the features (predictors)before training and inversely after predicting
- **scaler_y** – Scalar object that implements .transform and .inverse_transform, applied to the labels (targets)before training and inversely after predicting
- **printf** – print function to use for logging

Methods Summary

fit(fit_type, do_print)	fit all Model objects in collection
k_split(**kwargs)	apply hhpy.ds.k_split to self to create train-test or k-cross ready data
model_by_name(name, str])	extract a list of Models from the collection by their names
predict([X, df, return_type, ensemble, do_print])	predict with all models in collection
score(return_type, pivot, do_print, ...)	calculate score of the Model predictions
scoreplot([x, y, hue, hue_order, row, ...])	plot the score(s) using sns.barplot
train(df, k, groupby, str] = None, sortby, ...)	wrapper method that combined k_split, train, predict and score

Methods Documentation

fit(*fit_type: str = 'train_test'*, *do_print: bool = True*)

fit all Model objects in collection

Parameters

- **fit_type** – one of [‘train_test’, ‘k_cross’, ‘final’]
- **do_print** – Whether to print the steps to console

Returns None

k_split(***kwargs*)

apply hhpy.ds.k_split to self to create train-test or k-cross ready data

Parameters **kwargs** – keyword arguments passed to hhpy.ds.k_split

Returns None

model_by_name(*name: Union[list, str]*) → Union[hhpy.modelling.Model, list]
extract a list of Models from the collection by their names

Parameters **name** – name of the Model

Returns list of Models

predict(*X=None*, *df=None*, *return_type='self'*, *ensemble=False*, *do_print=True*)
predict with all models in collection

Parameters

- **X** – The feature (predictor) data used for predicting as DataFrame, np.array or column names
- **df** – Pandas DataFrame containing the predict data, optional if array like data is passed for X_predict
- **return_type** – one of ['y', 'df', 'DataFrame', 'self']
- **ensemble** – if True also predict with Ensemble like combinations of models. If True or mean calculatemean of individual predictions. If median calculate median of individual predictions.
- **do_print** – Whether to print the steps to console

Returns if return_type is self: None, else see Model.predict

score(*return_type: str = 'self'*, *pivot: bool = False*, *do_print: bool = True*, *display_score: bool = True*, ***kwargs*) → Optional[pandas.core.frame.DataFrame]
calculate score of the Model predictions

Parameters

- **return_type** – one of ['self', 'df', 'DataFrame']
- **pivot** – whether to pivot the DataFrame for easier readability [optional]
- **do_print** – Whether to print the steps to console
- **display_score** – Whether to display the score DataFrame
- **kwargs** – other keyword arguments passed to :func: *~hhpy.ds.df_score*

Returns if return_type is 'self': None, else: pandas DataFrame containing the scores

scoreplot(*x='y_ref', y='value', hue='model', hue_order=None, row='score', row_order=None, palette=['xkcd:blue', 'xkcd:red', 'xkcd:green', 'xkcd:cyan', 'xkcd:magenta', 'xkcd:golden yellow', 'xkcd:dark cyan', 'xkcd:red orange', 'xkcd:dark yellow', 'xkcd:easter green', 'xkcd:baby blue', 'xkcd:light brown', 'xkcd:strong pink', 'xkcd:light navy blue', 'xkcd:deep blue', 'xkcd:deep red', 'xkcd:ultramarine blue', 'xkcd:sea green', 'xkcd:plum', 'xkcd:old pink', 'xkcd:lawn green', 'xkcd:amber', 'xkcd:green blue', 'xkcd:yellow green', 'xkcd:dark mustard', 'xkcd:bright lime', 'xkcd:aquamarine', 'xkcd:very light blue', 'xkcd:light grey blue', 'xkcd:dark sage', 'xkcd:dark peach', 'xkcd:shocking pink'], width=16, height=4.5, scale=None, query=None, return_fig_ax=False, **kwargs*) → Optional[tuple]
plot the score(s) using sns.barplot

Parameters

- **x** – Name of the x variable in data or vector data
- **y** – Name of the y variable in data or vector data

- **hue** – Further split the plot by the levels of this variable [optional]
- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:

- `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
- `inv`: following sort of python standard sorting conventions but in inverse order
- `count`: sorted by value counts
- `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
- `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending

Parameters

- **row** – the variable to wrap around the rows [optional]
- **row_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:

- `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
- `inv`: following sort of python standard sorting conventions but in inverse order
- `count`: sorted by value counts
- `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
- `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending

Parameters

- **palette** – Collection of colors to be used for plotting. Can be a dictionary for with names for each level or a list of colors or an individual color name. Must be valid colors known to pyplot [optional]
- **width** – Width of each individual subplot [optional]
- **height** – Height of each individual subplot [optional]
- **scale** – scale the values [optional
- **query** – query to be passed to pd.DataFrame.query before plotting [optional]
- **return_fig_ax** – Whether to return the figure and axes objects as tuple to be captured as fig,ax = . . . , If False pyplot.show() is called and the plot returns None [optional]
- **kwargs** – other keyword arguments passed to sns.barplot

Returns see return_fig_ax

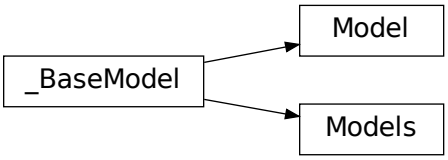
train(*df: pandas.core.frame.DataFrame = None, k: int = 5, groupby: Union[Sequence[T_co], str] = None, sortby: Union[Sequence[T_co], str] = None, random_state: int = None, fit_type: str = 'train_test', ensemble: bool = False, scores: Union[Sequence[T_co], str] = None, scale: float = None, do_predict: bool = True, do_score: bool = True, do_split: bool = True, do_fit: bool = True, do_print: bool = True, display_score: bool = True*)
wrapper method that combined k_split, train, predict and score

Parameters

- **df** – Pandas DataFrame containing the training and testing data. Can be saved to the Model object or supplied on an as needed basis.
- **k** – see hhpy.ds.k_split see hhpy.ds.k_split
- **groupby** – see hhpy.ds.k_split
- **sortby** – see hhpy.ds.k_split
- **random_state** – see hhpy.ds.k_split
- **fit_type** – see .fit
- **ensemble** – if True also predict with Ensemble like combinations of models. If True or mean calculatemean of individual predictions. If median calculate median of individual predictions.
- **scores** – see .score
- **scale** – see .score
- **do_print** – Whether to print the steps to console
- **display_score** – Whether to display the score DataFrame
- **do_split** – whether to apply k_split [optional]
- **do_fit** – whether to fit the Models [optional]
- **do_predict** – whether to add predictions to DataFrame [optional]
- **do_score** – whether to create self.df_score [optional]

Returns None

1.4.4 Class Inheritance Diagram



1.5 hhpy.plotting Module

1.5.1 hhpy.plotting.py

Contains plotting functions

1.5.2 Functions

heatmap(x, y, z, data, ax, cmap, agg_func, ...)	Wrapper for seaborn heatmap in x-y-z format
corrplot(data, annotations, number_format[, ax])	function to create a correlation plot using a seaborn heatmap based on: https://www.linkedin.com/pulse/generating-correlation-heatmaps-seaborn-python-andrew-holt
corrplot_bar(data, target, columns, ...)	correlation plot as barchart
pairwise_corrplot(data, corr_cutoff, ...)	print a pairwise_corrplot to for all variables in the df, by default only plots those with a correlation coefficient of >= corr_cutoff
distplot(x, str], data, hue, hue_order, ...)	Similar to seaborn.distplot but supports hues and some other things.
hist_2d(x, y, data, bins, std_cutoff, ...)	generic 2d histogram created by splitting the 2d area into equal sized cells, counting data points in them and drawn using pyplot.pcolormesh
paired_plot(data, cols, color, cmap, alpha, ...)	create a facet grid to analyze various aspects of correlation between two variables using seaborn.PairGrid
q_plim(s, q_min, q_max, offset_perc, ... [, ...])	returns quick x limits for plotting (cut off data not in q_min to q_max quantile)
levelplot(data, level, cols, str], hue, ...)	Plots a plot for each specified column for each level of a certain column plus a summary plot
get_legends(ax)	returns all legends on a given axis, useful if you have a secaxis
facet_wrap(func, data, facet, str], *args, ...)	modeled after r's facet_wrap function.
get_subax(ax, numpy.ndarray], row, col, ...)	shorthand to get around the fact that ax can be a 1D array or a 2D array (for subplots that can be 1x1,1xn,nx1)
ax_as_list(ax, numpy.ndarray])	takes any Axes and turns them into a list
ax_as_array(ax, numpy.ndarray])	takes any Axes and turns them into a numpy 2D array
rmsdplot(x, data, groups, str] = None, hue, ...)	creates a seaborn.barplot showing the rmsd calculating hhpy.ds.df_rmsd
insert_linebreak(s, pos, frac, max_breaks)	used to insert linebreaks in strings, useful for formatting axes labels
ax_tick_linebreaks(ax, x, y, **kwargs)	uses insert_linebreaks to insert linebreaks into the axes ticklabels
annotate_barplot(ax, x, y, ci, ci_newline, ...)	automatically annotates a barplot with bar values and error bars (if present).
animplot(data, x, y, t, lines, ...)	wrapper for FuncAnimation to be used with pandas DataFrames.
legend_outside(ax, width, loc, legend_space, ...)	draws a legend outside of the subplot
set_ax_sym(ax, x, y)	automatically sets the select axes to be symmetrical
custom_legend(colors, str], labels, str][, ...])	uses patches to create a custom legend with the specified colors
stemplot(x, y[, data, ax, color, baseline, ...])	modeled after pyplot.stemplot but more customizeable
get_twin(ax)	get the twin axis from an Axes object
get_axlim(ax, xy)	Wrapper function to get x limits, y limits or both with one function call
set_axlim(ax, lim, Mapping[KT, VT_co]], xy)	Wrapper function to set both x and y limits with one call
share_xy(ax, x, y, mode, adj_twin_ax)	set the subplots on the Axes to share x and/or y limits WITHOUT sharing x and y legends.
share_legend(ax, keep_i)	removes all legends except for i from an Axes object

Continued on next page

Table 8 – continued from previous page

barplot_err(x, y, xerr, yerr, data, **kwargs)	extension on seaborn barplot that allows for plotting errorbars with preprocessed data.The idea is based on this StackOverflow question.
countplot(x, str] = None, data, hue, ax, ...)	Based on seaborn barplot but with a few more options
quantile_plot(x, str], data, qs, ...)	plots the specified quantiles of a Series using seaborn.barplot

heatmap

hhpy.plotting.**heatmap**(*x: str*, *y: str*, *z: str*, *data: pandas.core.frame.DataFrame*, *ax: matplotlib.axes._axes.Axes = None*, *cmap: object = None*, *agg_func: str = 'mean'*, *invert_y: bool = True*, ***kwargs*) → matplotlib.axes._axes.Axes

Wrapper for seaborn heatmap in x-y-z format

Parameters

- **x** – Variable name for x axis value
- **y** – Variable name for y axis value
- **z** – Variable name for z value, used to color the heatmap
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **ax** – The axes object to plot on, defaults to current axis [optional]
- **cmap** – Color map to use [optional]
- **agg_func** – If more than one z value per x,y pair exists agg_func is used to aggregate the data. Must be a function name understood by pandas.DataFrame.agg
- **invert_y** – Whether to call ax.invert_yaxis (orders the heatmap as expected)
- **kwargs** – Other keyword arguments passed to seaborn heatmap

Returns The axes object with the plot on it

corrplot

hhpy.plotting.**corrplot**(*data: pandas.core.frame.DataFrame*, *annotations: bool = True*, *number_format: str = '.2f'*, *ax=None*)

function to create a correlation plot using a seaborn heatmap based on: https://www.linkedin.com/pulse/generating-correlation-heatmaps-seaborn-python-andrew-holt

Parameters

- **number_format** – The format string used for annotations [optional]
- **data** – Pandas DataFrame containing named data
- **annotations** – Whether to display annotations [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]

Returns The axes object with the plot on it

corrplot_bar

hhpy.plotting.**corrplot_bar**(*data: pandas.core.frame.DataFrame*, *target: str = None*, *columns: List[str] = None*, *corr_cutoff: float = 0*, *corr_as_alpha: bool = False*, *xlim: tuple = (-1, 1)*, *ax: matplotlib.axes._axes.Axes = None*)

correlation plot as barchart

Parameters

- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **target** – target variable name, if specified only correlations with the target are shown [optional]
- **columns** – columns for which to calculate the correlations, defaults to all numeric columns [optional]
- **corr_cutoff** – filter all correlation whose absolute value is below the cutoff [optional]
- **corr_as_alpha** – whether to set alpha value of bars to scale with correlation [optional]
- **xlim** – xlim scale for plot, defaults to (-1, 1) to show the absolute scale of the correlations. set to None if you want the plot x limits to scale to the highest correlation values [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]

Returns The axes object with the plot on it

pairwise_corrplot

hhpy.plotting.**pairwise_corrplot**(*data: pandas.core.frame.DataFrame, corr_cutoff: float = 0, col_wrap: int = 4, hue: str = None, hue_order: Union[list, str] = None, width: float = 7, height: float = 7, trendline: bool = True, alpha: float = 0.75, ax: matplotlib.axes._axes.Axes = None, target: str = None, palette: Union[Mapping[KT, VT_co], Sequence[T_co], str] = ['xkcd:blue', 'xkcd:red', 'xkcd:green', 'xkcd:cyan', 'xkcd:magenta', 'xkcd:golden yellow', 'xkcd:dark cyan', 'xkcd:red orange', 'xkcd:dark yellow', 'xkcd:easter green', 'xkcd:baby blue', 'xkcd:light brown', 'xkcd:strong pink', 'xkcd:light navy blue', 'xkcd:deep blue', 'xkcd:deep red', 'xkcd:ultramarine blue', 'xkcd:sea green', 'xkcd:plum', 'xkcd:old pink', 'xkcd:lawn green', 'xkcd:amber', 'xkcd:green blue', 'xkcd:yellow green', 'xkcd:dark mustard', 'xkcd:bright lime', 'xkcd:aquamarine', 'xkcd:very light blue', 'xkcd:light grey blue', 'xkcd:dark sage', 'xkcd:dark peach', 'xkcd:shocking pink'], max_n: int = 10000, random_state: int = None, sample_warn: bool = True, return_fig_ax: bool = True, **kwargs*) → Optional[tuple]

print a pairwise_corrplot to for all variables in the df, by default only plots those with a correlation coefficient of >= corr_cutoff

Parameters

- **data** – Pandas DataFrame containing named data
- **corr_cutoff** – filter all correlation whose absolute value is below the cutoff [optional]
- **col_wrap** – After how many columns to create a new line of subplots [optional]
- **hue** – Further split the plot by the levels of this variable [optional]

- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **width** – Width of each individual subplot [optional]
- **height** – Height of each individual subplot [optional]
- **trendline** – Whether to add a trendline [optional]
- **alpha** – Alpha transparency level [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]
- **target** – target variable name, if specified only correlations with the target are shown [optional]
- **palette** – Collection of colors to be used for plotting. Can be a dictionary for with names for each level or a list of colors or an individual color name. Must be valid colors known to pyplot [optional]
- **max_n** – Maximum number of samples to be used for plotting, if this number is exceeded max_n samples are drawnat random from the data which triggers a warning unless sample_warn is set to False.Set to False or None to use all samples for plotting. [optional]
- **random_state** – Random state (seed) used for drawing the random samples [optional]
- **sample_warn** – Whether to trigger a warning if the data has more samples than max_n [optional]
- **return_fig_ax** – Whether to return the figure and axes objects as tuple to be captured as fig,ax = . . . , If False pyplot.show() is called and the plot returns None [optional]
- **kwargs** – other keyword arguments passed to pyplot.subplots

Returns if return_fig_ax: figure and axis objects as tuple, else None

distplot

hhpy.plotting.**distplot**(*x: Union[Sequence[T_co], str], data: pandas.core.frame.DataFrame = None, hue: str = None, hue_order: Union[Sequence[T_co], str] = 'sorted', palette: Union[Mapping[KT, VT_co], Sequence[T_co], str] = None, linecolor: str = 'black', edgecolor: str = 'black', alpha: float = None, bins: Union[Sequence[T_co], int] = 40, perc: bool = None, top_nr: int = None, other_name: str = 'other', title: bool = True, title_prefix: str = '', std_cutoff: float = None, hist: bool = None, distfit: Union[str, bool, None] = 'kde', fill: bool = True, legend: bool = True, legend_loc: str = None, legend_space: float = 0.1, legend_ncol: int = 1, agg_func: str = 'mean', number_format: str = '.2f', kde_steps: int = 1000, max_n: int = 100000, random_state: int = None, sample_warn: bool = True, xlim: Sequence[T_co] = None, linestyle: str = None, label_style: str = 'mu_sigma', x_offset_perc: float = 0.025, ax: matplotlib.axes._axes.Axes = None, **kwargs*) → matplotlib.axes._axes.Axes

Similar to seaborn.distplot but supports hues and some other things. Plots a combination of a histogram and a kernel density estimation.

Parameters

- **x** – the name of the variable(s) in data or vector data, if data is provided and x is a list of columns the DataFrame is automatically melted and the newly generated column used as hue. i.e. you plot the distributions of multiple columns on the same axis
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **hue** – Further split the plot by the levels of this variable [optional]
- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **palette** – Collection of colors to be used for plotting. Can be a dictionary for with names for each level or a list of colors or an individual color name. Must be valid colors known to pyplot [optional]
- **linecolor** – Color of the kde fit line, overwritten with palette by hue level if hue is specified [optional]
- **edgecolor** – Color of the histogram edges [optional]
- **alpha** – Alpha transparency level [optional]
- **bins** – Nr of bins of the histogram [optional]
- **perc** – Whether to display the y-axes as percentage, if false count is displayed. Defaults if hue: True, else False [optional]
- **top_nr** – limit hue to top_nr levels using hhpy.ds.top_n, the rest will be cast to other [optional]

- **other_name** – name of the other group created by hhpy.ds.top_n [optional]
- **title** – whether to set the plot title equal to x's name [optional]
- **title_prefix** – prefix to be used in plot title [optional]
- **std_cutoff** – automatically cutoff data outside of the std_cutoff standard deviations range, by default this is off but a recommended value for a good visual experience without outliers is 3 [optional]
- **hist** – whether to show the histogram, default False if hue else True [optional]
- **distfit** – one of ['kde', 'gauss', 'False', 'None']. If 'kde' fits a kernel density distribution to the data. If gauss fits a gaussian distribution with the observed mean and std to the data. [optional]
- **fill** – whether to fill the area under the distfit curve, ignored if hist is True [optional]
- **legend** – Whether to show a legend [optional]
- **legend_loc** – Location of the legend, one of [bottom, right] or accepted value of pyplot.legendIf in [bottom, right] legend_outside is used, else pyplot.legend [optional]
- **legend_space** – Only valid if legend_loc is bottom. The space between the main plot and the legend [optional]
- **legend_ncol** – Number of columns to use in legend [optional]
- **agg_func** – one of ['mean', 'median']. The agg function used to find the center of the distribution [optional]
- **number_format** – The format string used for annotations [optional]
- **kde_steps** – Nr of steps the range is split into for kde fitting [optional]
- **max_n** – Maximum number of samples to be used for plotting, if this number is exceeded max_n samples are drawnat random from the data which triggers a warning unless sample_warn is set to False.Set to False or None to use all samples for plotting. [optional]
- **random_state** – Random state (seed) used for drawing the random samples [optional]
- **sample_warn** – Whether to trigger a warning if the data has more samples than max_n [optional]
- **xlim** – X limits for the axis as tuple, passed to ax.set_xlim() [optional]
- **linestyle** – Linestyle used, must a valid linestyle recognized by pyplot.plot [optional]
- **label_style** – one of ['mu_sigma', 'plain']. If mu_sigma then the mean (or median) and std value are displayed inside the label [optional]
- **x_offset_perc** – the amount whitespace to display next to x_min and x_max in percent of x_range [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]
- **kwargs** – additional keyword arguments passed to pyplot.plot

Returns The axes object with the plot on it

hist_2d

hhpy.plotting.**hist_2d**(*x: str*, *y: str*, *data: pandas.core.frame.DataFrame*, *bins: int = 100*, *std_cutoff: int = 3*, *cutoff_perc: float = 0.01*, *cutoff_abs: float = 0*, *cmap: str = 'rainbow'*, *ax: matplotlib.axes._axes.Axes = None*, *color_sigma: str = 'xkcd:red'*, *draw_sigma: bool = True*, ***kwargs*) → matplotlib.axes._axes.Axes

generic 2d histogram created by splitting the 2d area into equal sized cells, counting data points in them and drawn using pyplot.pcolormesh

Parameters

- **x** – Name of the x variable in data or vector data
- **y** – Name of the y variable in data or vector data
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **bins** – Nr of bins of the histogram [optional]
- **std_cutoff** – remove data outside of std_cutoff standard deviations, for a good visual experience try 3 [optional]
- **cutoff_perc** – if less than this percentage of data points is in the cell then the data is ignored [optional]
- **cutoff_abs** – if less than this amount of data points is in the cell then the data is ignored [optional]
- **cmap** – Color map to use [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]
- **color_sigma** – color to highlight the sigma range in, must be a valid pyplot.plot color [optional]
- **draw_sigma** – whether to highlight the sigma range [optional]
- **kwargs** – other keyword arguments passed to pyplot.pcolormesh [optional]

Returns The axes object with the plot on it

paired_plot

hhpy.plotting.**paired_plot**(*data: pandas.core.frame.DataFrame*, *cols: Sequence[T_co]*, *color: str = None*, *cmap: str = None*, *alpha: float = 1*, ***kwargs*) → seaborn.axisgrid.FacetGrid

create a facet grid to analyze various aspects of correlation between two variables using seaborn.PairGrid

Parameters

- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **cols** – list of exactly two variables to be compared
- **color** – color used for plotting, must be known to matplotlib.pyplot [optional]
- **cmap** – Color map to use [optional]
- **alpha** – Alpha transparency level [optional]
- **kwargs** – other arguments passed to seaborn.PairGrid

Returns seaborn FacetGrid object with the plots on it

q_plim

hhpy.plotting.**q_plim**(*s: pandas.core.series.Series*, *q_min: float = 0.1*, *q_max: float = 0.9*, *offset_perc: float = 0.1*, *limit_min_max: bool = False*, *offset=True*) → tuple

returns quick x limits for plotting (cut off data not in q_min to q_max quantile)

Parameters

- **s** – pandas Series to truncate
- **q_min** – lower bound quantile [optional]
- **q_max** – upper bound quantile [optional]
- **offset_perc** – percentage of offset to the left and right of the quantile boundaries
- **limit_min_max** – whether to truncate the plot limits at the data limits
- **offset** – whether to apply the offset

Returns a tuple containing the x limits

levelplot

hhpy.plotting.**levelplot**(*data: pandas.core.frame.DataFrame*, *level: str*, *cols: Union[list, str]*, *hue: str = None*, *order: Union[list, str] = None*, *hue_order: Union[list, str] = None*, *func: Callable = <function distplot>*, *summary_title: bool = True*, *level_title: bool = True*, *do_print: bool = False*, *width: int = 7*, *height: int = 7*, *return_fig_ax: bool = True*, *kwargs_subplots_adjust: Mapping[KT, VT_co] = None*, *kwargs_summary: Mapping[KT, VT_co] = None*, ***kwargs*) → Union[None, tuple]

Plots a plot for each specified column for each level of a certain column plus a summary plot

Parameters

- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **level** – the name of the column to split the plots by, must be in data
- **cols** – the columns to create plots for, defaults to all numeric columns [optional]
- **hue** – Further split the plot by the levels of this variable [optional]
- **order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)

- `inv`: following sort of python standard sorting conventions but in inverse order
- `count`: sorted by value counts
- `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
- `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending

- **`func`** – function to use for plotting, must support 1 positional argument, data, hue, ax and kwargs [optional]
- **`summary_title`** – whether to automatically set the summary plot title [optional]
- **`level_title`** – whether to automatically set the level plot title [optional]
- **`do_print`** – whether to print intermediate steps to console [optional]
- **`width`** – Width of each individual subplot [optional]
- **`height`** – Height of each individual subplot [optional]
- **`return_fig_ax`** – Whether to return the figure and axes objects as tuple to be captured as fig,ax = . . . , If False pyplot.show() is called and the plot returns None [optional]
- **`kwargs_subplots_adjust`** – other keyword arguments passed to pyplot.subplots_adjust [optional]
- **`kwargs_summary`** – other keyword arguments passed to summary distplot, if None uses kwargs [optional]
- **`kwargs`** – other keyword arguments passed to func [optional]

Returns see return_fig_ax

get_legends

hhpy.plotting.**`get_legends`**(*ax: matplotlib.axes._axes.Axes = None*) → list

returns all legends on a given axis, useful if you have a secaxis

Parameters **`ax`** – The axes object to plot on, defaults to current axis [optional]

Returns list of legends

facet_wrap

hhpy.plotting.**`facet_wrap`**(*func: Callable, data: pandas.core.frame.DataFrame, facet: Union[list, str], *args, facet_type: str = None, col_wrap: int = 4, width: int = 7, height: int = 7, catch_error: bool = True, return_fig_ax: bool = True, sharex: bool = False, sharey: bool = False, show_xlabel: bool = True, x_tick_rotation: int = None, y_tick_rotation: int = None, ax_title: str = 'set', order: Union[list, str] = None, subplots_kws: Mapping[KT, VT_co] = None, **kwargs*)

modeled after r's facet_wrap function. Wraps a number of subplots onto a 2d grid of subplots while creating a new line after col_wrap columns. Uses a given plot function and creates a new plot for each facet level.

Parameters

- **`func`** – Any plot function. Must support keyword arguments data and ax
- **`data`** – Pandas DataFrame containing named data, optional if vector data is used

- **facet** – The column / list of columns to facet over.
- **args** – passed to func
- **facet_type** – one of [‘group’, ‘cols’, None]. If group facet is treated as the column creating the facet levels and a subplot is created for each level. If cols each facet is in turn passed as the first positional argument to the plot function func. If None then the facet_type is inferred: a single facet value will be treated as group and multiple facet values will be treated as cols.
- **col_wrap** – After how many columns to create a new line of subplots [optional]
- **width** – Width of each individual subplot [optional]
- **height** – Height of each individual subplot [optional]
- **catch_error** – whether to keep going in case of an error being encountered in the plot function [optional]
- **return_fig_ax** – Whether to return the figure and axes objects as tuple to be captured as fig,ax = . . . , If False pyplot.show() is called and the plot returns None [optional]
- **sharex** – Whether to share the x axis [optional]
- **sharey** – Whether to share the y axis [optional]
- **show_xlabel** – whether to show the x label for each subplot
- **x_tick_rotation** – x tick rotation for each subplot
- **y_tick_rotation** – y tick rotation for each subplot
- **ax_title** – one of [‘set’,’hide’], if set sets axis title to facet name, if hide forcefully hids axis title
- **order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **subplots_kws** – other keyword arguments passed to pyplot.subplots
- **kwargs** – other keyword arguments passed to func

Returns if return_fig_ax: figure and axis objects as tuple, else None

examples

Check out the example notebook

get_subax

hhpy.plotting.**get_subax**(*ax: Union[matplotlib.axes._axes.Axes, numpy.ndarray], row: int = None, col: int = None, rows_prio: bool = True*) → matplotlib.axes._axes.Axes

shorthand to get around the fact that ax can be a 1D array or a 2D array (for subplots that can be 1x1,1xn,nx1)

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **row** – Row index [optional]
- **col** – Column index [optional]
- **rows_prio** – decides if to use row or col in case of a 1xn / nx1 shape (False means cols get priority)

Returns The axes object with the plot on it

ax_as_list

hhpy.plotting.**ax_as_list**(*ax: Union[matplotlib.axes._axes.Axes, numpy.ndarray]*) → list

takes any Axes and turns them into a list

Parameters **ax** – The axes object to plot on, defaults to current axis [optional]

Returns List containing the subaxes

ax_as_array

hhpy.plotting.**ax_as_array**(*ax: Union[matplotlib.axes._axes.Axes, numpy.ndarray]*) → numpy.ndarray

takes any Axes and turns them into a numpy 2D array

Parameters **ax** – The axes object to plot on, defaults to current axis [optional]

Returns Numpy 2D array containing the subaxes

rmsdplot

hhpy.plotting.**rmsdplot**(*x: str, data: pandas.core.frame.DataFrame, groups: Union[Sequence[T_co], str] = None, hue: str = None, hue_order: Union[Sequence[T_co], str] = None, cutoff: float = 0, ax: matplotlib.axes._axes.Axes = None, color_as_balance: bool = False, balance_cutoff: float = None, rmsd_as_alpha: bool = False, sort_by_hue: bool = False, palette=None, barh_kws=None, **kwargs*)

creates a seaborn.barplot showing the rmsd calculating hhpy.ds.df_rmsd

Parameters

- **x** – Name of the x variable in data or vector data
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **groups** – the columns to calculate the rmsd for, defaults to all columns [optional]
- **hue** – Further split the plot by the levels of this variable [optional]
- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:

- `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
- `inv`: following sort of python standard sorting conventions but in inverse order
- `count`: sorted by value counts
- `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
- `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending

- **`cutoff`** – drop rmsd values smaller than cutoff [optional]
- **`ax`** – The axes object to plot on, defaults to current axis [optional]
- **`color_as_balance`** – whether to color the bars based on how balanced the levels are [optional]
- **`balance_cutoff`** – if specified the balance coloring red for worse balance than balance cutoff [optional]
- **`rmsd_as_alpha`** – whether to use set the alpha values of the columns based on the rmsd value [optional]
- **`sort_by_hue`** – passed to hhpy.ds.df_rmsd [optional]
- **`palette`** – Collection of colors to be used for plotting. Can be a dictionary for with names for each level or a list of colors or an individual color name. Must be valid colors known to pyplot [optional]
- **`barh_kws`** – other keyword arguments passed to seaborn.barplot [optional]
- **`kwargs`** – other keyword arguments passed to hhpy.ds.rf_rmsd [optional]

Returns The axes object with the plot on it

insert_linebreak

hhpy.plotting.**`insert_linebreak`**(*s: str*, *pos: int = None*, *frac: float = None*, *max_breaks: int = None*) → str

used to insert linebreaks in strings, useful for formatting axes labels

Parameters

- **`s`** – string to insert linebreaks into
- **`pos`** – inserts a linebreak every pos characters [optional]
- **`frac`** – inserts a linebreak after frac percent of characters [optional]
- **`max_breaks`** – maximum number of linebreaks to insert [optional]

Returns string with the linebreaks inserted

ax_tick_linebreaks

hhpy.plotting.**`ax_tick_linebreaks`**(*ax: matplotlib.axes._axes.Axes = None*, *x: bool = True*, *y: bool = True*, ***kwargs*) → None

uses insert_linebreaks to insert linebreaks into the axes ticklabels

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **x** – whether to insert linebreaks into the x axis label [optional]
- **y** – whether to insert linebreaks into the y axis label [optional]
- **kwargs** – other keyword arguments passed to insert_linebreaks

Returns None

annotate_barplot

hhpy.plotting.**annotate_barplot**(*ax: matplotlib.axes._axes.Axes = None*, *x: Sequence[T_co] = None*, *y: Sequence[T_co] = None*, *ci: bool = True*, *ci_newline: bool = True*, *adj_ylim: float = 0.05*, *nr_format: str = '.2f'*, *ha: str = 'center'*, *va: str = 'center'*, *offset: int = 10.0*, ***kwargs*) → matplotlib.axes._axes.Axes

automatically annotates a barplot with bar values and error bars (if present). Currently does not work with ticks!

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **x** – Name of the x variable in data or vector data
- **y** – Name of the y variable in data or vector data
- **ci** – whether to annotate error bars [optional]
- **ci_newline** – whether to add a newline between values and error bar values [optional]
- **adj_ylim** – whether to automatically adjust the plot y limits to fit the annotations [optional]
- **nr_format** – The format string used for annotations [optional]
- **ha** – horizontal alignment [optional]
- **va** – vertical alignment [optional]
- **offset** – offset between bar top and annotation center [optional]
- **kwargs** – other keyword arguments passed to pyplot.annotate

Returns The axes object with the plot on it

animplot

hhpy.plotting.**animplot**(*data: pandas.core.frame.DataFrame = None*, *x: str = 'x'*, *y: str = 'y'*, *t: str = 't'*, *lines: Mapping[KT, VT_co] = None*, *max_interval: int = None*, *time_per_frame: int = 200*, *mode: str = 'jshtml'*, *title: bool = True*, *title_prefix: str = ''*, *t_format: str = None*, *fig: matplotlib.figure.Figure = None*, *ax: matplotlib.axes._axes.Axes = None*, *color: str = None*, *label: str = None*, *legend: bool = False*, *legend_out: bool = False*, *legend_kws: Mapping[KT, VT_co] = None*, *xlim: tuple = None*, *ylim: tuple = None*, *ax_facecolor: Union[str, Mapping[KT, VT_co]] = None*, *grid: bool = False*, *vline: Union[Sequence[T_co], float] = None*, ***kwargs*) → Union[IPython.core.display.HTML, matplotlib.animation.FuncAnimation]

wrapper for FuncAnimation to be used with pandas DataFrames. Assumes that you have a DataFrame containing one data point for each x-y-t combination.

If mode is set to jshtml the function is optimized for use with Jupyter Notebook and returns an Interactive JavaScript Widget.

Parameters

- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **x** – Name of the x variable in data
- **y** – Name of the y variable in data
- **t** – Name of the t variable in data
- **lines** – you can also pass lines that you want to animate. Details to follow [optional]
- **max_interval** – max interval at which to abort the animation [optional]
- **time_per_frame** – time per frame [optional]
- **mode** – one of the below [optional]
  - `matplotlib`: Return the matplotlib FuncAnimation object
  - `html`: Returns an HTML5 movie (You need to install ffmpeg for this to work)
  - `jshtml`: Returns an interactive Javascript Widget
- **title** – whether to set the time as plot title [optional]
- **title_prefix** – title prefix to be put in front of the time if title is true [optional]
- **t_format** – format string used to format the time variable in the title [optional]
- **fig** – figure to plot on [optional]
- **ax** – axes to plot on [optional]
- **color** – color used for plotting, must be known to matplotlib.pyplot [optional]
- **label** – label to use for the data [optional]
- **legend** – Whether to show a legend [optional]
- **legend_out** – Whether to draw the legend outside of the axis, can also be a location string [optional]
- **legend_kws** – Other keyword arguments passed to pyplot.legend [optional]
- **xlim** – X limits for the axis as tuple, passed to ax.set_xlim() [optional]
- **ylim** – Y limits for the axis as tuple, passed to ax.set_ylim() [optional]
- **ax_facecolor** – passed to ax.set_facecolor, can also be a conditional mapping to change the facecolor at specific timepoints t [optional]
- **grid** – Whether to toggle ax.grid() [optional]
- **vline** – A list of x positions to draw vlines at [optional]
- **kwargs** – other keyword arguments passed to pyplot.plot

Returns see mode

examples

Check out the example notebook

legend_outside

hhpy.plotting.**legend_outside**(*ax: matplotlib.axes._axes.Axes = None*, *width: float = 0.85*, *loc: str = 'right'*, *legend_space: float = 0.1*, *offset_x: float = 0*, *offset_y: float = 0*, ***kwargs*)

draws a legend outside of the subplot

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **width** – how far to shrink down the subplot if loc=='right'
- **loc** – one of ['right','bottom'], where to put the legend
- **legend_space** – how far below the subplot to put the legend if loc=='bottom'
- **offset_x** – x offset for the legend
- **offset_y** – y offset for the legend
- **kwargs** – other keyword arguments passed to pyplot.legend

Returns None

set_ax_sym

hhpy.plotting.**set_ax_sym**(*ax: matplotlib.axes._axes.Axes*, *x: bool = True*, *y: bool = True*)

automatically sets the select axes to be symmetrical

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **x** – whether to set x axis to be symmetrical
- **y** – whether to set y axis to be symmetrical

Returns None

custom_legend

hhpy.plotting.**custom_legend**(*colors: Union[list, str]*, *labels: Union[list, str]*, *do_show=True*) → Optional[list]

uses patches to create a custom legend with the specified colors

Parameters

- **colors** – list of matplotlib colors to use for the legend
- **labels** – list of labels to use for the legend
- **do_show** – whether to show the created legend

Returns if do_show: None, else handles

stemplot

hhpy.plotting.**stemplot**(*x*, *y*, *data=None*, *ax=None*, *color='xkcd:blue'*, *baseline=0*, *kwline=None*, ***kwargs*)

modeled after pyplot.stemplot but more customizeable

Parameters

- **x** – Name of the x variable in data or vector data
- **y** – Name of the y variable in data or vector data
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **ax** – The axes object to plot on, defaults to current axis [optional]
- **color** – color used for plotting, must be known to matplotlib.pyplot [optional]
- **baseline** – where to draw the baseline for the stemplot
- **kwline** – other keyword arguments passed to pyplot.plot
- **kwargs** – other keyword arguments passed to pyplot.scatter

Returns The axes object with the plot on it

get_twin

hhpy.plotting.**get_twin**(*ax: matplotlib.axes._axes.Axes*) → Optional[matplotlib.axes._axes.Axes]
get the twin axis from an Axes object

Parameters **ax** – The axes object to plot on, defaults to current axis [optional]

Returns the twin axis if it exists, else None

get_axlim

hhpy.plotting.**get_axlim**(*ax: matplotlib.axes._axes.Axes*, *xy: Optional[str] = None*) → Union[tuple, Mapping[KT, VT_co]]
Wrapper function to get x limits, y limits or both with one function call

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **xy** – one of [‘x’, ‘y’, ‘xy’, None]

Returns if xy is ‘xy’ or None returns a dictionary else returns x or y lims as tuple

set_axlim

hhpy.plotting.**set_axlim**(*ax: matplotlib.axes._axes.Axes, lim: Union[Sequence[T_co], Mapping[KT, VT_co]], xy: Optional[str] = None*)
Wrapper function to set both x and y limits with one call

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **lim** – axes limits as tuple or Mapping
- **xy** – one of [‘x’, ‘y’, ‘xy’, None]

Returns None

share_xy

hhpy.plotting.**share_xy**(*ax: matplotlib.axes._axes.Axes*, *x: bool = True*, *y: bool = True*, *mode: str = 'all'*, *adj_twin_ax: bool = True*)

set the subplots on the Axes to share x and/or y limits WITHOUT sharing x and y legends. If you want that please use pyplot.subplots(share_x=True,share_y=True) when creating the plots.

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **x** – whether to share x limits [optional]
- **y** – whether to share y limits [optional]
- **mode** – one of ['all', 'row', 'col'], if all shares across all subplots, else just across rows / columns
- **adj_twin_ax** – whether to also adjust twin axes

Returns None

share_legend

hhpy.plotting.**share_legend**(*ax: matplotlib.axes._axes.Axes*, *keep_i: int = None*)

removes all legends except for i from an Axes object

Parameters

- **ax** – The axes object to plot on, defaults to current axis [optional]
- **keep_i** – index of the plot whose legend you want to keep

Returns None

barplot_err

hhpy.plotting.**barplot_err**(*x: str*, *y: str*, *xerr: str = None*, *yerr: str = None*, *data: pandas.core.frame.DataFrame = None*, ***kwargs*) → matplotlib.axes._axes.Axes

extension on seaborn barplot that allows for plotting errorbars with preprocessed data. The idea is based on this StackOverflow question

Parameters

- **x** – Name of the x variable in data
- **y** – Name of the y variable in data
- **xerr** – variable to use as x error bars [optional]
- **yerr** – variable to use as y error bars [optional]
- **data** – Pandas DataFrame containing named data
- **kwargs** – other keyword arguments passed to seaborn barplot

Returns The axes object with the plot on it

countplot

hhpy.plotting.**countplot**(*x: Union[Sequence[T_co], str] = None, data: pandas.core.frame.DataFrame = None, hue: str = None, ax: matplotlib.axes._axes.Axes = None, order: Union[Sequence[T_co], str] = None, hue_order: Union[Sequence[T_co], str] = None, normalize_x: bool = False, normalize_hue: bool = False, palette: Union[Mapping[KT, VT_co], Sequence[T_co], str] = None, x_tick_rotation: int = None, count_twinx: bool = False, hide_legend: bool = False, annotate: bool = True, annotate_format: str = ',.0f', legend_kws: Mapping[KT, VT_co] = None, barplot_kws: Mapping[KT, VT_co] = None, count_twinx_kws: Mapping[KT, VT_co] = None, **kwargs*)

Based on seaborn barplot but with a few more options

Parameters

- **x** – Name of the x variable in data or vector data
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **hue** – Further split the plot by the levels of this variable [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]
- **order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **normalize_x** – whether to normalize x, causes the sum of each x group to be 100 percent [optional]
- **normalize_hue** – whether to normalize hue, causes the sum of each hue group to be 100 percent [optional]

- **palette** – Collection of colors to be used for plotting. Can be a dictionary for with names for each level or a list of colors or an individual color name. Must be valid colors known to pyplot [optional]
- **x_tick_rotation** – Set x tick label rotation to this value [optional]
- **count_twinx** – whether to plot the count values on the second axis (if using normalize) [optional]
- **hide_legend** – whether to hide the legend [optional]
- **annotate** – whether to use annotate_barplot [optional]
- **annotate_format** – The format string used for annotations [optional]
- **legend_kws** – additional keyword arguments passed to pyplot.legend [optional]
- **barplot_kws** – additional keyword arguments passed to seaborn.barplot [optional]
- **count_twinx_kws** – additional keyword arguments passed to pyplot.plot [optional]
- **kwargs** – additional keyword arguments passed to hhpy.ds.df_count [optional]

Returns The axes object with the plot on it

quantile_plot

hhpy.plotting.**quantile_plot**(*x: Union[Sequence[T_co], str], data: pandas.core.frame.DataFrame = None, qs: Union[Sequence[T_co], float] = None, x2: str = None, hue: str = None, hue_order: Union[Sequence[T_co], str] = None, to_abs: bool = False, ax: matplotlib.axes._axes.Axes = None, **kwargs*) → matplotlib.axes._axes.Axes

plots the specified quantiles of a Series using seaborn.barplot

Parameters

- **x** – Name of the x variable in data or vector data
- **data** – Pandas DataFrame containing named data, optional if vector data is used
- **qs** – Quantile levels [optional]
- **x2** – if specified: subtracts x2 from x before calculating quantiles [optional]
- **hue** – Further split the plot by the levels of this variable [optional]
- **hue_order** – Either a string describing how the (hue) levels or to be ordered or an explicit list of levels to beused for plotting. Accepted strings are:
  - `sorted`: following python standard sorting conventions (alphabetical for string, ascending for value)
  - `inv`: following sort of python standard sorting conventions but in inverse order
  - `count`: sorted by value counts
  - `mean`, `mean_ascending`, `mean_descending`: sorted by mean value, defaults to descending
  - `median`, `mean_ascending`, `median_descending`: sorted by median value, defaults to descending
- **to_abs** – whether to cast the values to absolute before proceeding [optional]
- **ax** – The axes object to plot on, defaults to current axis [optional]

- **kwargs** – other keyword arguments passed to seaborn.barplot

Returns The axes object with the plot on it

CHAPTER 2

quickstart

Try these functions to get a feeling of what the package does

Main

Quick convenience

- *tprint*
- *is_list_like*
- *force_list*

DS

Quick ways to filter, score and split DataFrames

- *qf*
- *f_score*
- *k_split*
- *df_split*

Modelling

Modelling provides an extension on any statistical model that implements .fit and .predict that allows for easy multi modelling.

- *Model*
- *Models*

Plotting

Animated plots and distribution plots

- *animplot*
- *distplot*
- *facet_wrap*

CHAPTER 3

Indices and tables

- genindex
- modindex
- search

Python Module Index

h

Index