



Sample Scientist

Bob Brederveld
Luke Moth

Introduction

- system analyses: datasets
- often performing same analyses again
- until now: often re-script or make same graphs in excel

More efficient using tools: *Sample Scientist*

Platform for performing standard analytics

EcoConverter – Dataprofeet – Sample Scientist

Explanation tool

- works on all “n samples vs m observables”
- supports (for now) the following analyses:
 - PCA
 - hierarchical clustering
 - linear regression
 - stepwise regression
 - correlation matrix
 - boxplots (including ANOVA test when subgroups)
 - Stepwise with PCs
 - IR EGV plot
- available as python package or as stand-alone executable

dummy data

some_index	some_location	some_timestamp	some_variable1	some_variable2	some.
0	my_location1	2021-03-21	5.0	99.9	0.000
1	my_location2	2021-04-21	-5.0	44.9	0.000

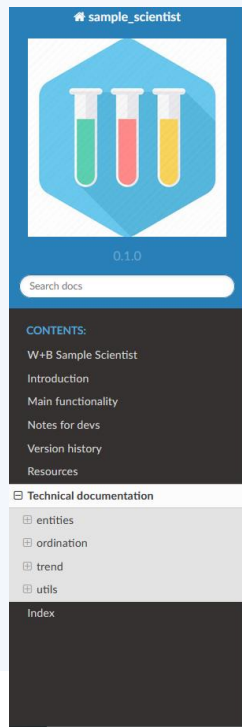
Explanation tool - input

- When using the exe version:
a *_config.yaml

```

1  ---
2  data input:
3  ...# REQUIRED
4  ...data_file: "D:/repos/trendanalyse/tests/data/demo_data_bluecan.csv"
5  ...data_file_type: "csv" ...# may be csv, xlsx or xlsx-multitab
6  ...# optional, defaults which user can update
7  ...csv_seperator: ";"
8  ...decimal_seperator: ",",
9  ...# IF metadata cols and/or respons cols in output, please specify
10 ...# otherwise these will be interpreted like numeric (and set to nan)
11 cols:
12 ...cols_metadata:
13 ...- "datum"
14 ...- "locatie"
15 ...- "type_weer"
16 ...cols_respons:
17 ...- "total_ghg"
18 ...col_datetime: "datum"
19 ...cols_color_def: ["locatie", "type_weer"]
20 ...cols_ignore: ["data_mist"]
21 ...# OPTIONAL (delete if need be)
22 ...output_dir: "D:/other_work/tests_eco_analyser/testexe6"
23 pca: $delete_entire_PCA_block_if_no_PCA_is_required
24 ...# OPTIONAL
25 ...include_varimax: true
26 ...varimax_min_axes: 2
27 ...varimax_max_axes: 8
28 ...biplot_max_axes: 10
29 linear_regression: $delete_entire_linear_regression_block_if_no_linear_re
30 ...# OPTIONAL
  
```

Explanation tool – use in python



box_plot_anova.py

module that contains BoxPlotAnova class

Use this class when you have a dataframe with n samples and m observables and one (or multiple) 'meta' observables to group data. The boxplots (and/or anova and/or post hoc tukey tests) will then show how these groups differ (and whether this is likely to be significant).

```
class sample_scientist.entities.box_plot_anova.BoxPlotAnova(df_data, cols_groupby=None,
cols_obs=None, export_dir=None) [source]
```

init class with dataframe. Dataframe must have columns as provided in list_cols_group (and other observables)

Parameters: `df_data : pd.DataFrame`

`cols_groupby : list, optional`

if not provided very 'boring plots' of single boxplot in axis if provided data will be grouped by these cols, hence creating n groups for each observable. Each group will be single 'box' in boxplot

```
classmethod from_CorrelationData(my_data, cols_groupby, export_dir) [source]
```

init class from instance of CorrelationData - provide data discard columns metadata

```
get_data_groups(col_obs) [source]
```

get data for each group

```
make_boxplot(col_obs, include_anova_p=True) [source]
```

make boxplot for single obs

```
make_boxplots(include_anova_p=True) [source]
```

make boxplots for all obs

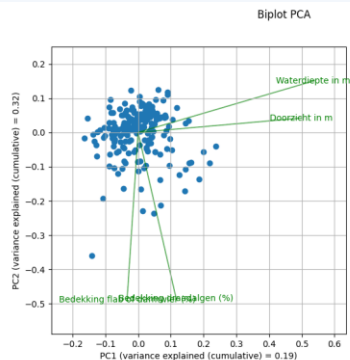
```
one_way_anova(col_obs) [source]
```

perform one way anova

Explanation tool - output

- Folder structure
- Plots (color coded)
- Excel files
- logging

Name	Date modified	Type
boxplots	20-4-2022 14:25	File folc
cluster	20-4-2022 14:21	File folc
cor_matrix	20-4-2022 14:25	File folc
lin_reg	20-4-2022 14:21	File folc
log	20-4-2022 14:20	File folc
pca	20-4-2022 14:20	File folc
pv_cluster	20-4-2022 14:25	File folc
step	20-4-2022 14:21	File folc
step_pc	20-4-2022 14:25	File folc



20220420-142016.log - Notepad

```

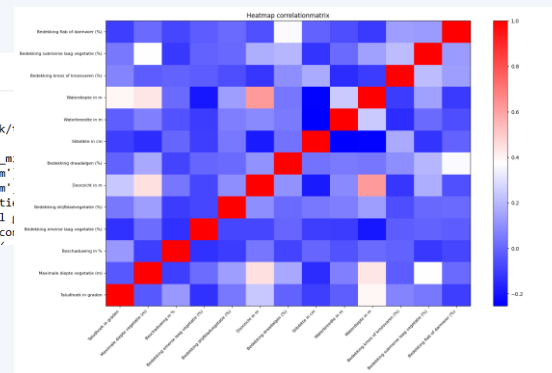
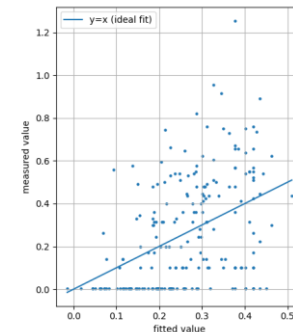
File Edit Format View Help
[2022-04-20 14:20:16]; [INFO] ]; _init_ -> Checking *config.yaml file
[2022-04-20 14:20:16]; [INFO] ]; main -> Running tool with version 0.1.0
[2022-04-20 14:20:16]; [INFO] ]; main -> Results will be saved in D:/other_work/
[2022-04-20 14:20:16]; [INFO] ]; provide_correlation_data -> Reading data
[2022-04-20 14:20:16]; [INFO] ]; _check_cols_existence -> Using columns ['data_m
[2022-04-20 14:20:16]; [INFO] ]; _check_cols_existence -> Using columns ['datum'
[2022-04-20 14:20:16]; [INFO] ]; _check_cols_existence -> Using columns ['datum'
[2022-04-20 14:20:16]; [INFO] ]; _check_cols_existence -> Using columns ['locati
[2022-04-20 14:20:16]; [INFO] ]; _check_cols_existence -> Using columns ['total
[2022-04-20 14:20:16]; [WARNING] ]; _check_datatypes_columns -> Column data_mist cor

```

```

n_kreeften_totaal =
0.0 -
0.0log(beschaduwung_in_percent) -
0.0log(bedekking_submerse_laag_vegetatie_percent) +
0.2sqrt(arscin(maximale_diepte_vegetatie_m/2)) -
0.0log(bedekking_flab_of_darmwier_percent)
r2=0.17

```





www.witteveenbos.com