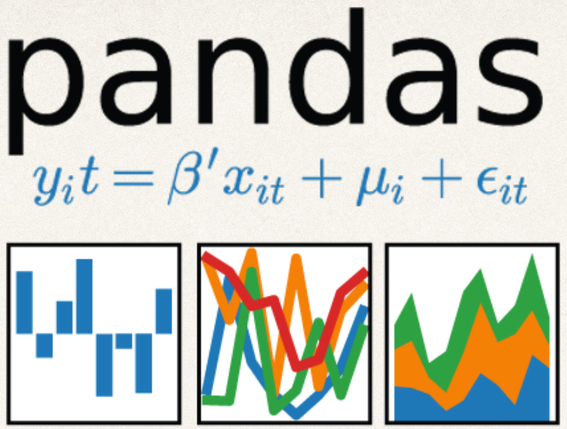
DataFrameを Python数行で ＥＤＡ

malo21st



EDA（探索的データ解析）
- データを料理する前に、どのようなデータが与えられているか確認すること。
- この段階を踏むことで、データに対する理解が深まる。
- EDAには、データの集計、要約、可視化が含まれる。

第10回 意思決定のためのデータ分析勉強会 online　2020/06/20

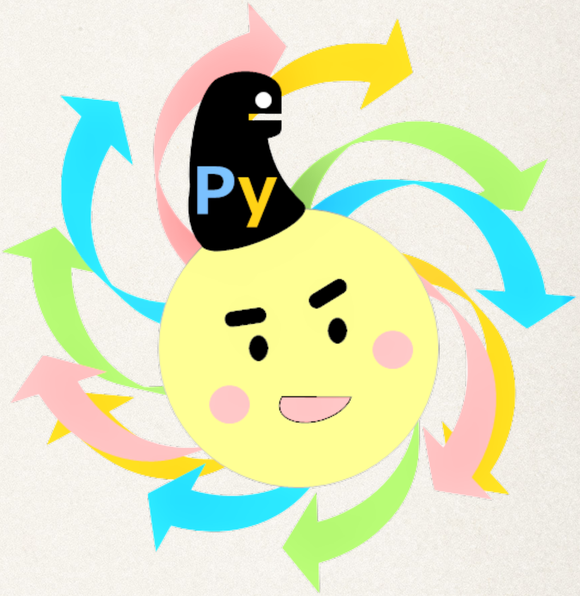
～ リモート環境でも使える意思決定術！ ～

おまえだれよ

- 田中丸　祐治（たなかまる　ゆうじ）

  malo21st（まろツウェンティーファースト）

- ＩＴ業界とは関係のない完全趣味でPythonやってます。
- 福岡を中心に、Pythonコミュニティに出没しています。
- Github ： https://github.com/malo21st
- Twitter ： @malo21st



なぜ、malo21st なの？

malo は、高校の国語の先生が
「田中麻呂」と呼んだのが始まり。
21st は、今が 21世紀 だから。

本日の流れ

機械学習でお馴染みの アヤメのDataFrame

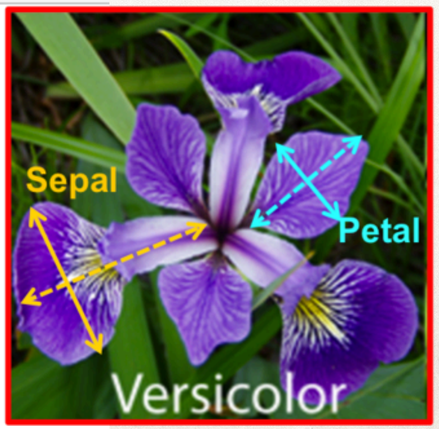
を

Python 数行で いろいろＥＤＡ

```python
from sklearn import datasets
import pandas as pd

iris = datasets.load_iris()
df_iris = pd.DataFrame(iris.data, columns=iris.feature_names)
df_iris['target'] = iris.target_names[iris.target]
df_iris.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



何はともあれ、DataFrame.describe( )

```
1 df_iris.describe()
```

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

pandas_profiling.ProfileReport(DataFrame)

```
1  import pandas_profiling as pdp  # pip install pandas-profiling
2
3  pdp.ProfileReport(df_iris)
```

Overview

Dataset info

Number of variables	5
Number of observations	150
Total Missing (%)	0.0%
Total size in memory	5.9 KiB
Average record size in memory	40.5 B

Variables types

Numeric	3
Categorical	1
Boolean	0
Date	0
Text (Unique)	0
Rejected	1
Unsupported	0

Warnings

- petal_width is highly correlated with petal_length (ρ = 0.96287) **Rejected**
- Dataset has 1 duplicate rows **Warning**

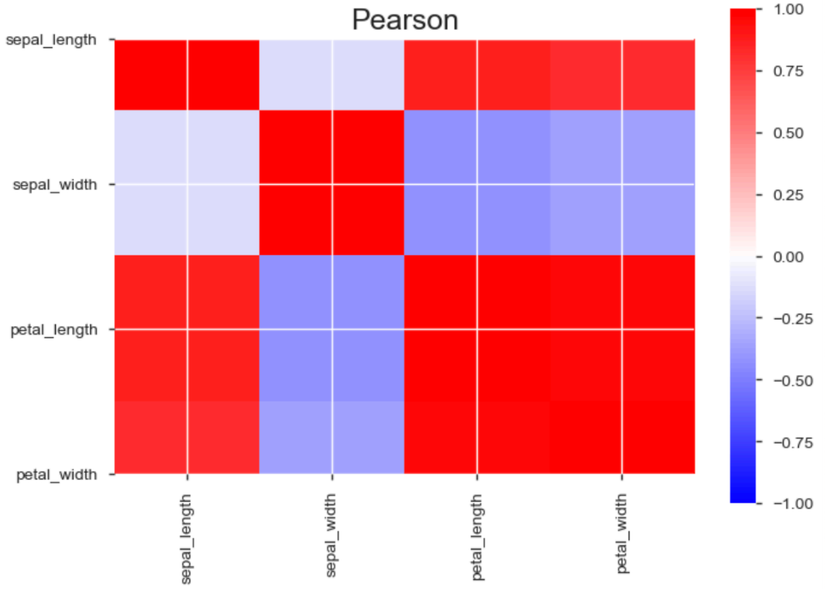
Variables

petal_length
Numeric

Distinct count	43	Mean	3.758
Unique (%)	28.7%	Minimum	1
Missing (%)	0.0%	Maximum	6.9
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

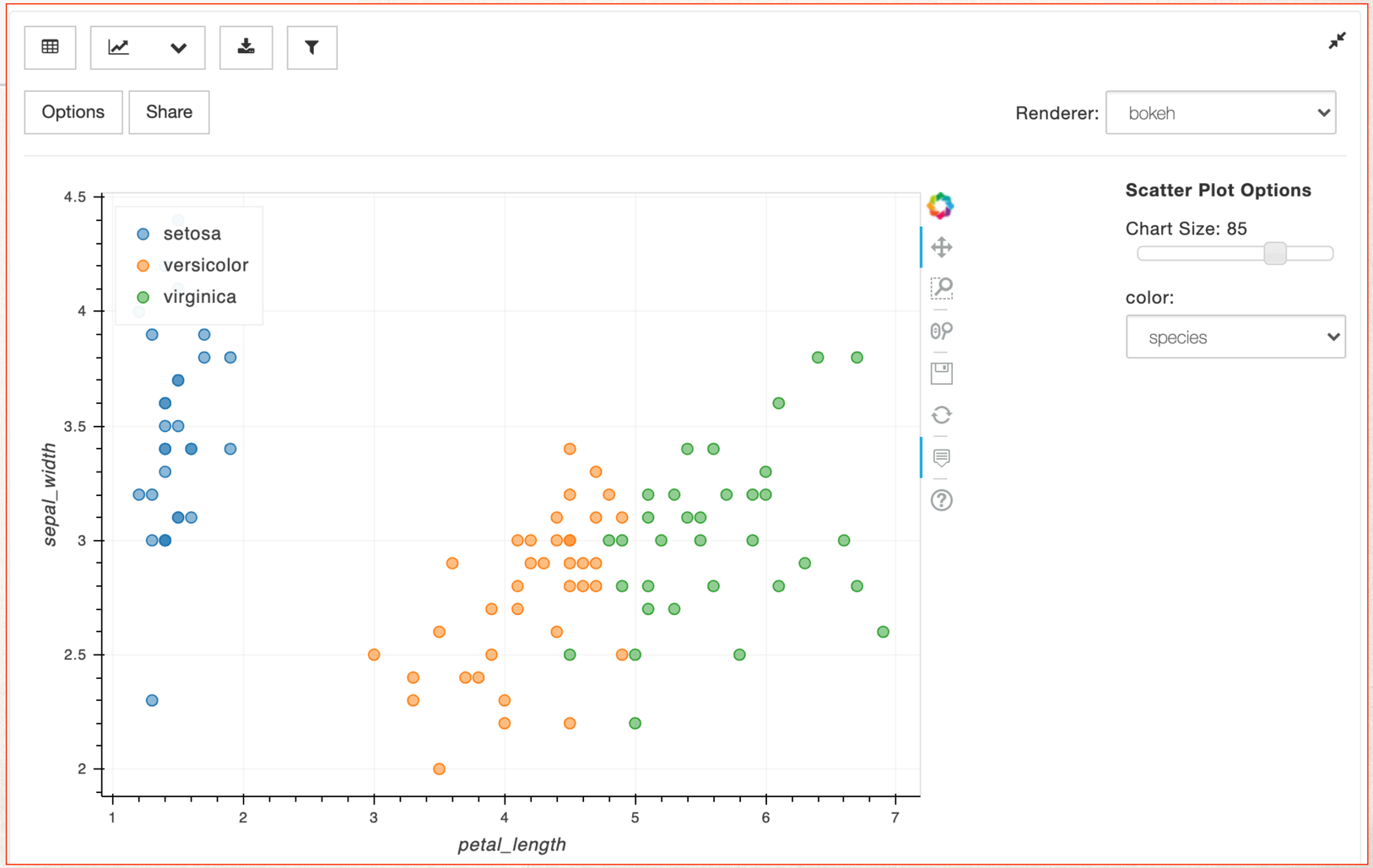
Toggle details

Correlations



pixiedust display(DataFrame)

```
1  import pixiedust  # pip install pixiedust
2
3  display(df_iris)
```

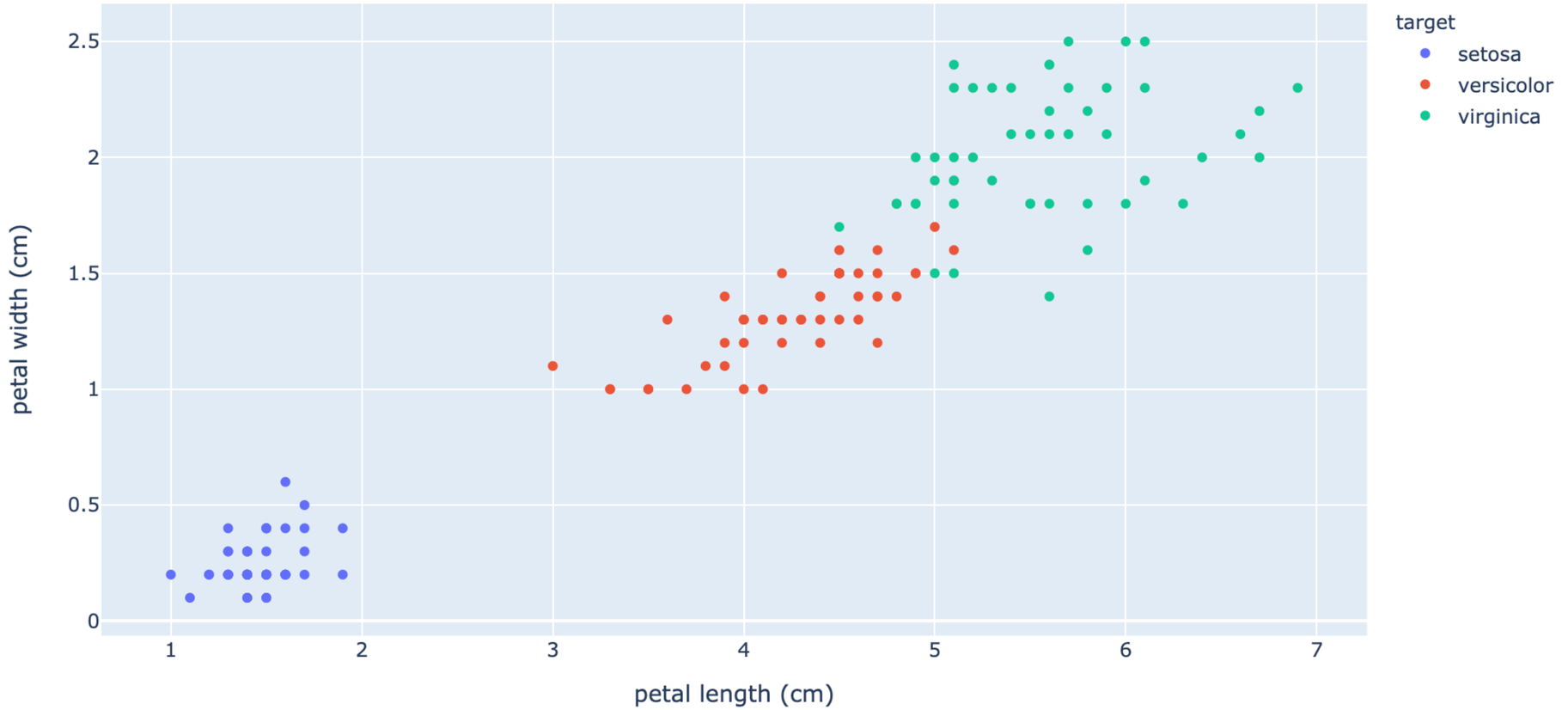


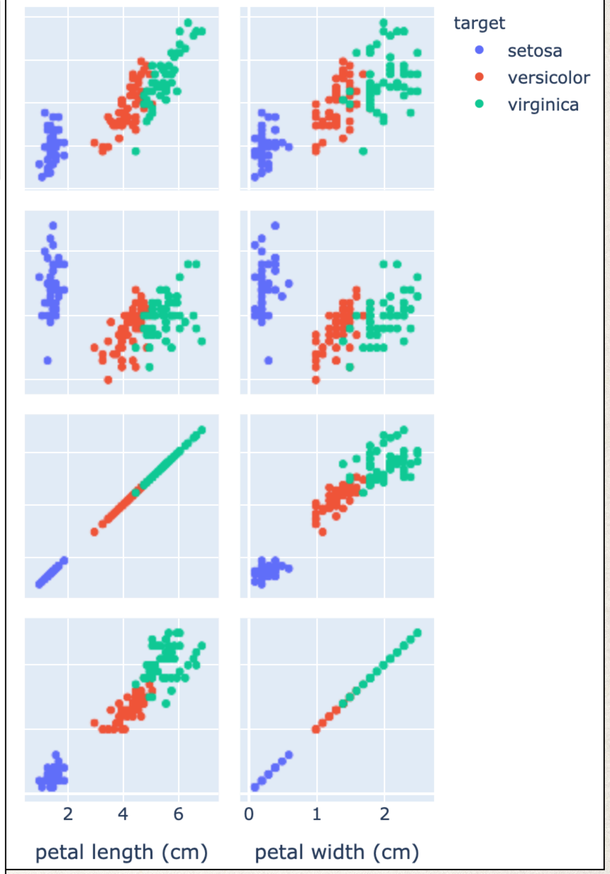
plotly.express

```
fig = px.scatter_matrix(df_iris, color='target', width=800, height=700,
                        dimensions=['sepal length (cm)', 'sepal width (cm)',
                                    'petal length (cm)','petal width (cm)'])
fig
```

```
import plotly.express as px

fig = px.scatter(df_iris, x='petal length (cm)', y='petal width (cm)', color='target')
fig
```



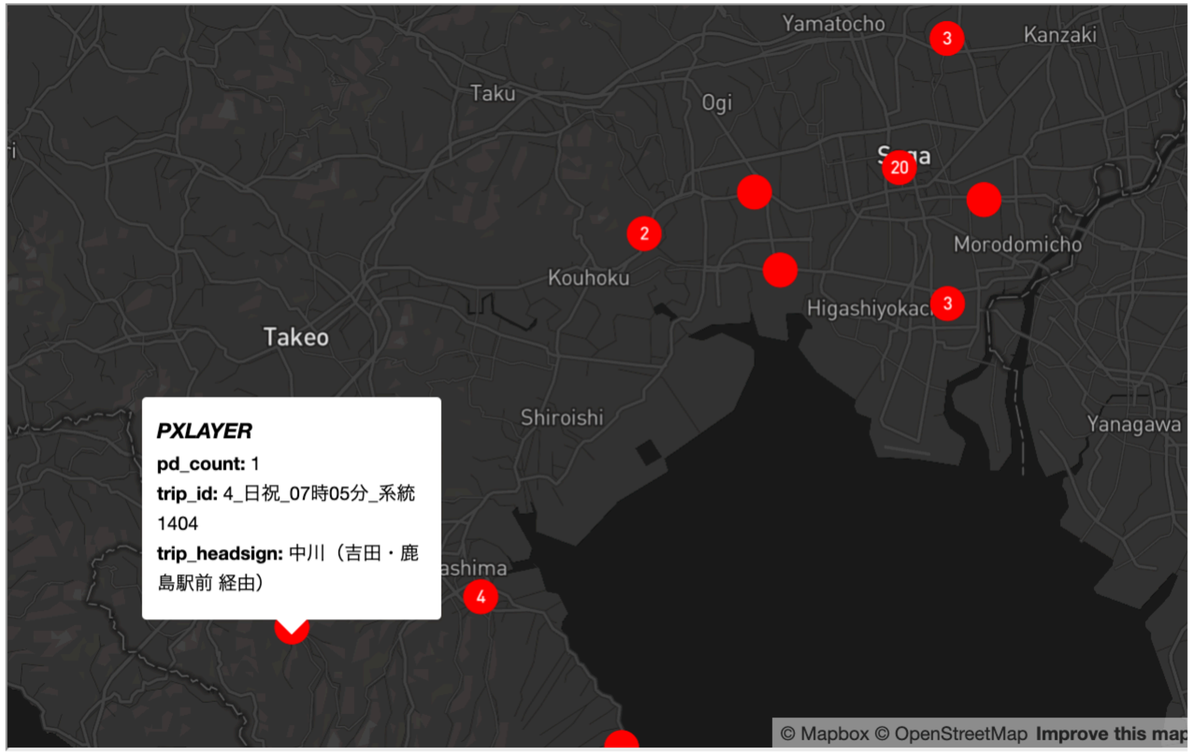


pixiedust　display(DataFrame)　緯度・経度

```
1 import pixiedust
2
3 display(df_bus)
```

Options　Share　Renderer: mapbox

佐賀のバスの現在位置



Map View Options
Show legend
Map Size: 100
Opacity: 100
Style: simple-cluster
Color Ramp: Light to Dark Red
basemap: dark-v9
Number of Bins: 8
Bin Options

ご清聴 ありがとうございました。

- 本日の資料

https://github.com/malo21st/DAD4D200620

第10回 意思決定のためのデータ分析勉強会 online　　2020/06/20

～ リモート環境でも使える意思決定術！ ～

（補足）関連サイト

- pandas https://pandas.pydata.org/
- pandas_profiling https://github.com/pandas-profiling/pandas-profiling
- pixiedust https://github.com/pixiedust/pixiedust
- Plotly Express https://plotly.com/python/plotly-express/