Warsztaty modelowania

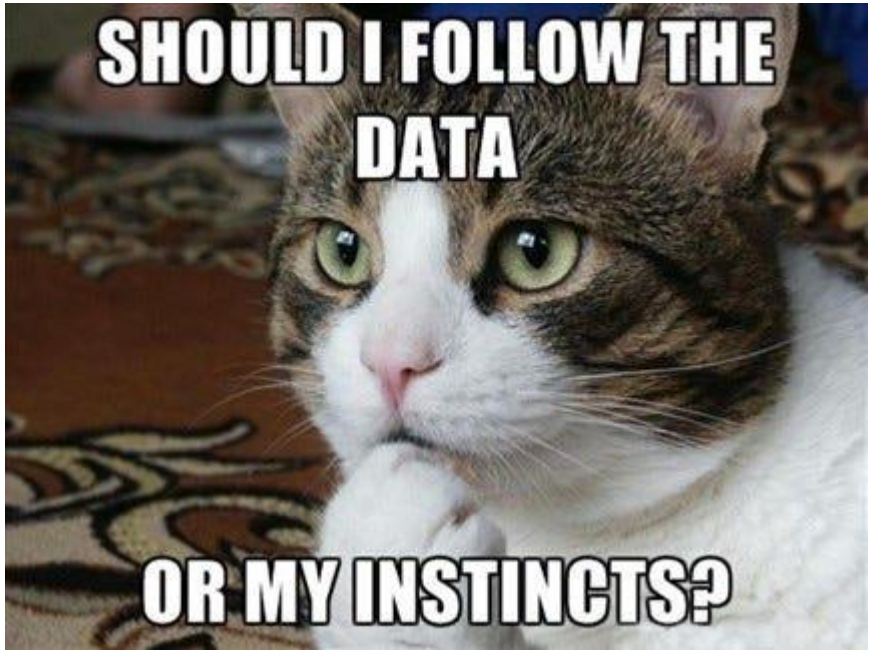
02 – Analiza eksploracyjna danych (EDA)

opracowała Patrycja Naumczyk

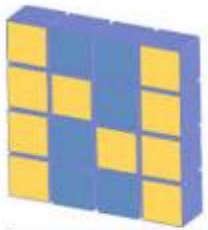
EDA - Exploratory Data Analysis

Absolutne minimum

1. Typy danych
2. Braki danych
3. Duplikaty
4. Statystyki opisowe (adekwatne do skali pomiarowej zmiennej)
5. Rozkłady zmiennych
6. Zależności między zmiennymi



Typy danych



NumPy

python™

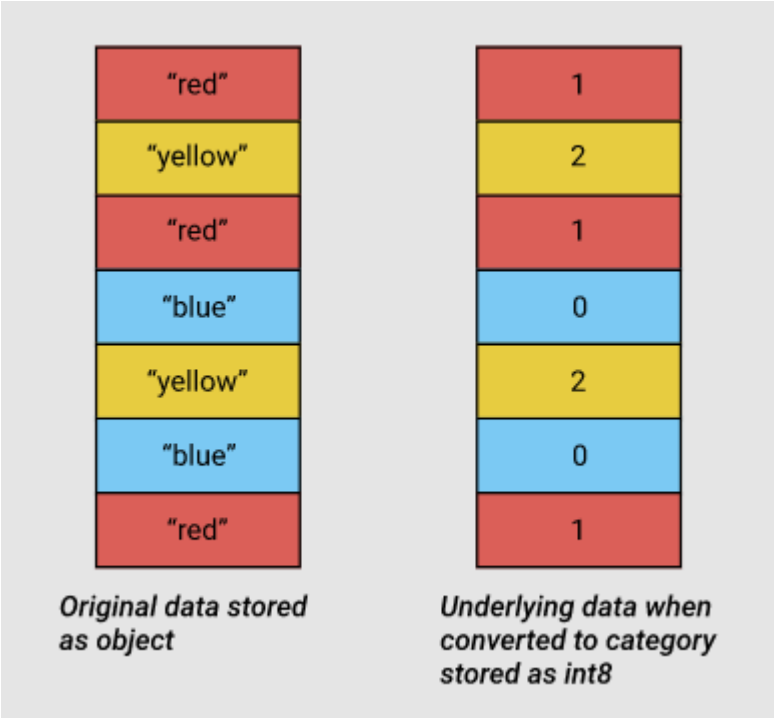
Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

Kind of Data	pandas Data Type	Scalar	Array
TZ-aware datetime	`DatetimeTZDtype`	`Timestamp`	Datetimes
Timedeltas	(none)	`Timedelta`	Timedeltas
Period (time spans)	`PeriodDtype`	`Period`	Periods
Intervals	`IntervalDtype`	`Interval`	Intervals
Nullable Integer	`Int64Dtype`, ...	(none)	Nullable integer
Nullable Float	`Float64Dtype`, ...	(none)	Nullable float
Categorical	`CategoricalDtype`	(none)	Categoricals
Sparse	`SparseDtype`	(none)	Sparse
Strings	`StringDtype`	`str`	Strings
Nullable Boolean	`BooleanDtype`	`bool`	Nullable Boolean
PyArrow	`ArrowDtype`	Python Scalars or `NA`	PyArrow

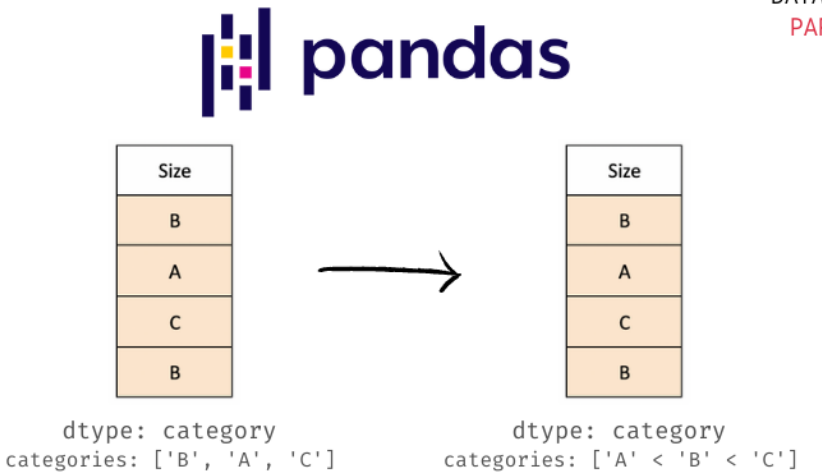
https://pandas.pydata.org/docs/reference/arrays.html
https://pandas.pydata.org/docs/user_guide/basics.html#basics-dtypes

Typy danych

1. Atrybut .dtypes / .dtype
2. Metody
   a) astype()
   b) to_numeric()
   c) to_datetime()
   d) select_dtypes()



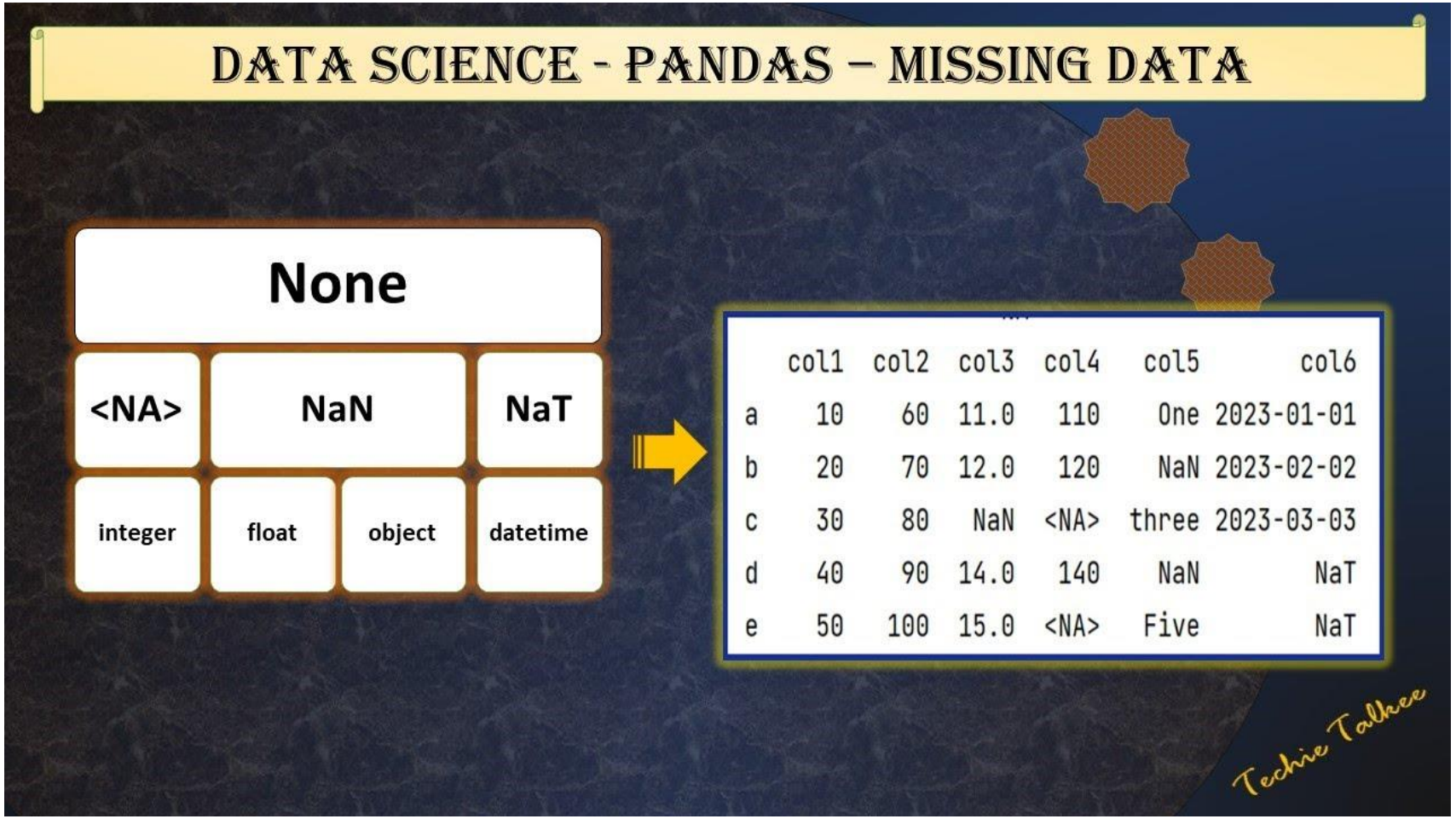
DATA SCIENCE
PARICHAY



Set Category Order

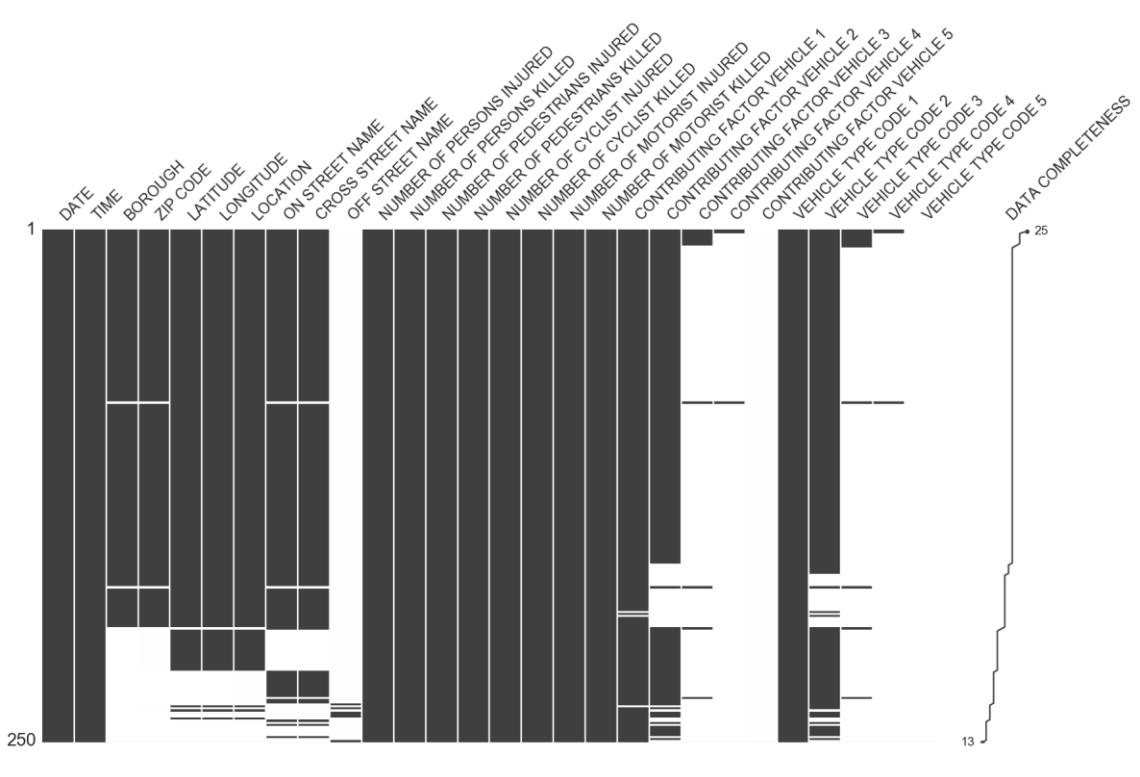
Braki danych

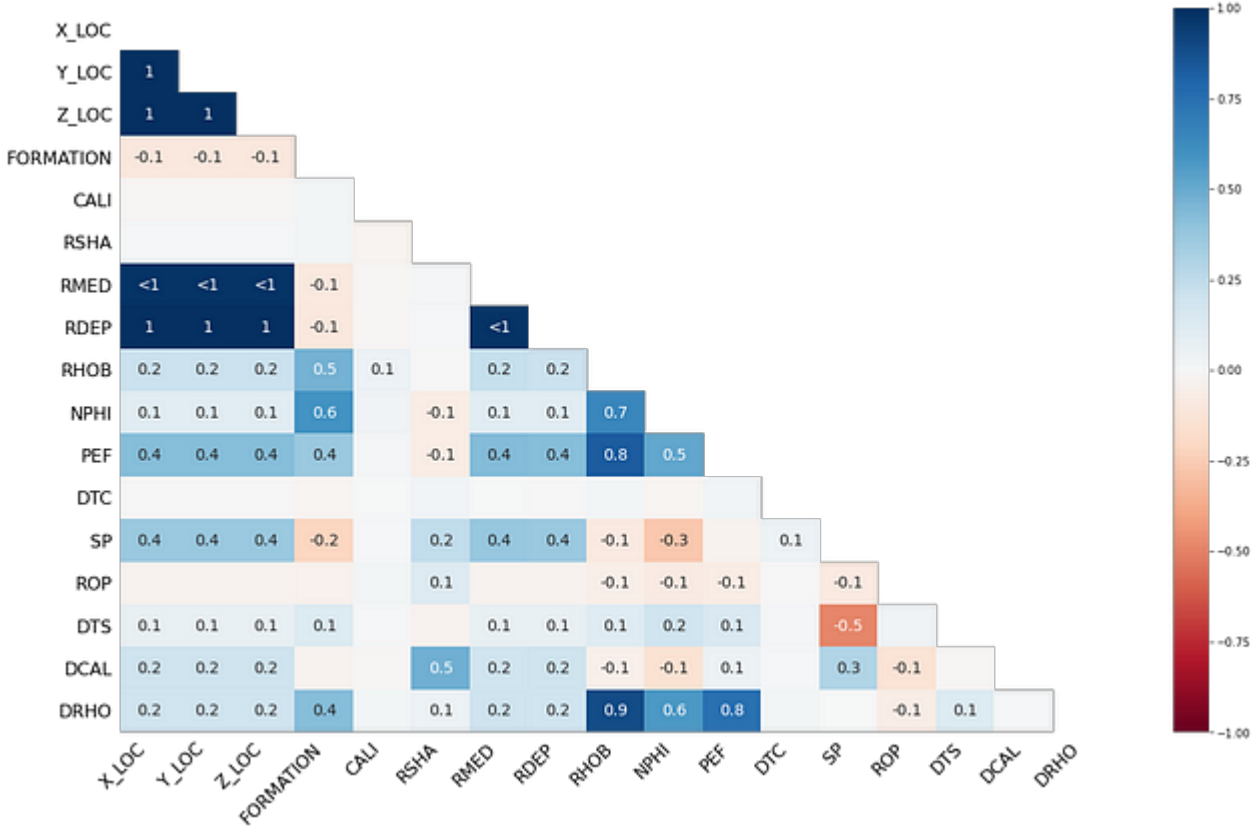
1. Reprezentacja braku
   a) np.nan
   b) NaT
   c) NA – string i nullable
   d) None
2. Metody
   a) isna() / isnull()
   b) notna()
   c) dropna()
   d) fillna() / interpolate()



https://pandas.pydata.org/docs/user_guide/missing_data.html

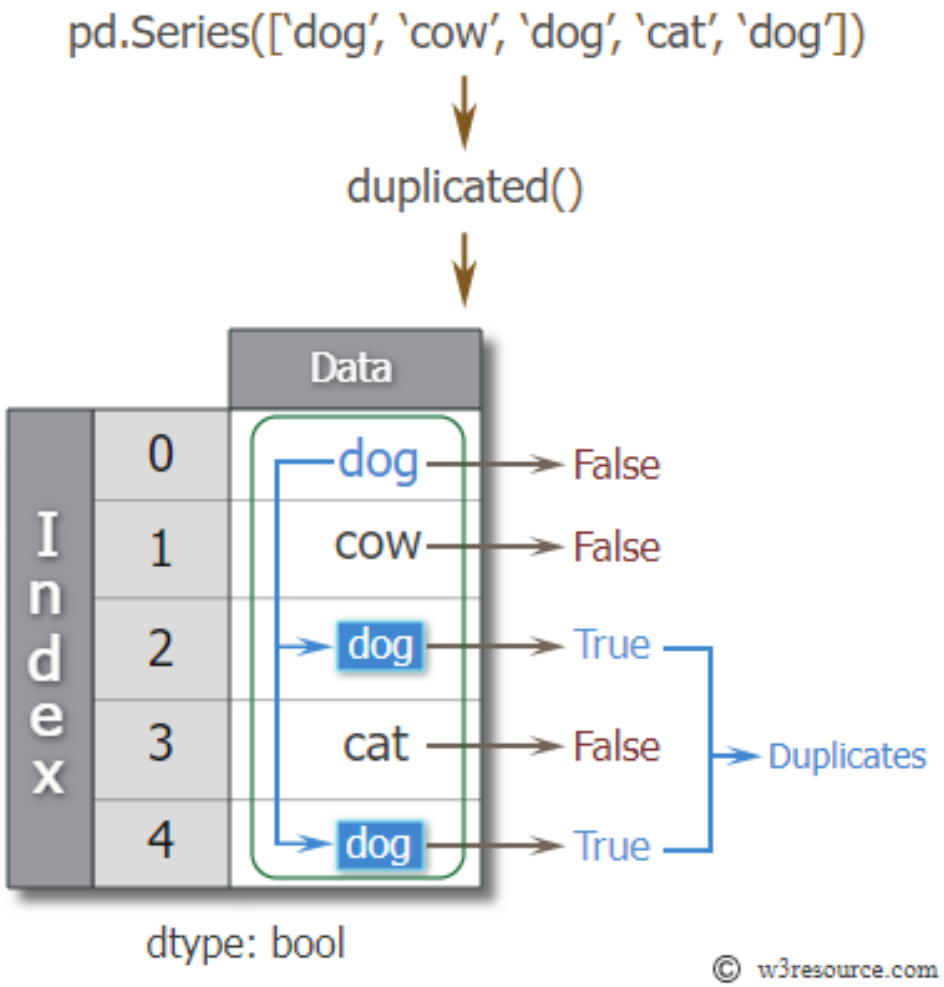
Braki danych - missingno





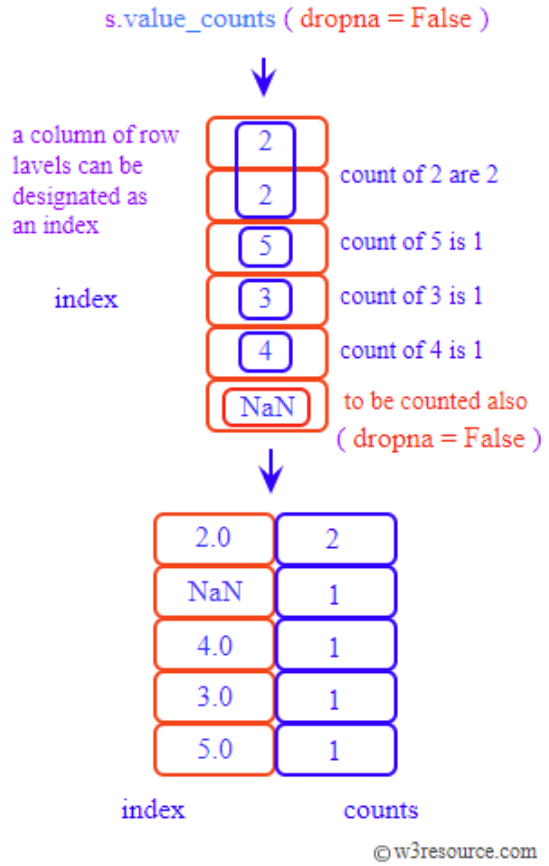
Duplikaty danych

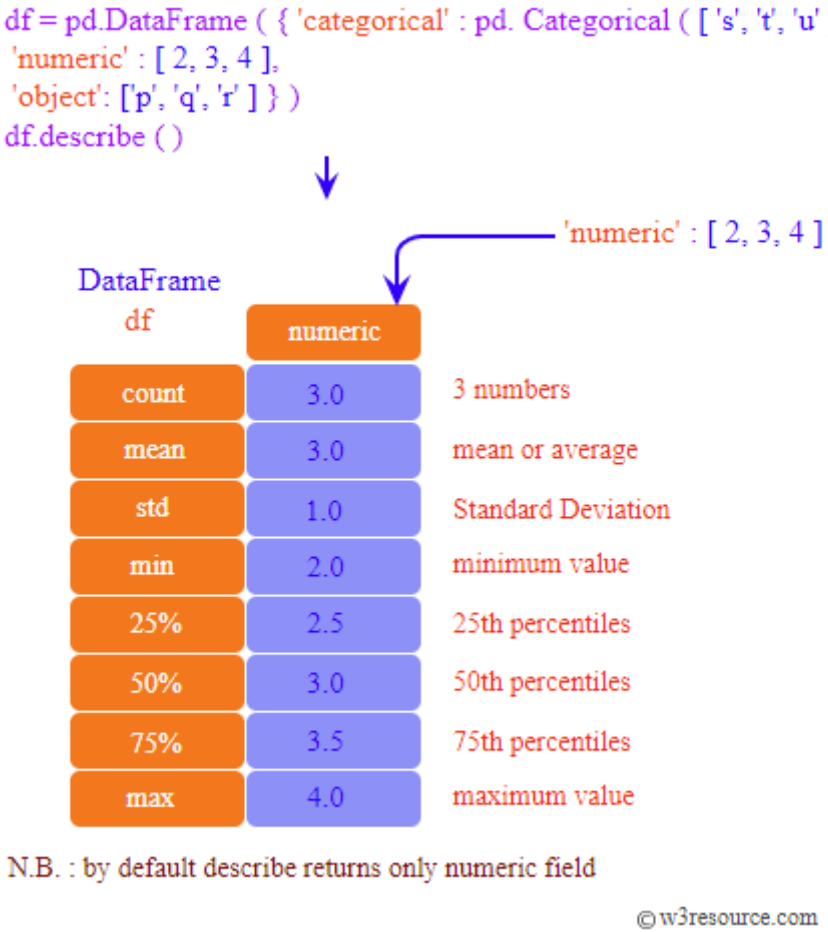
1. Flagi metod konstruowania obiektów
2. Metody badania indexu
   a) .index.is_unique
   b) .index.duplicated()
3. Metody badania treści
   a) duplicated()
   b) drop_duplicates()

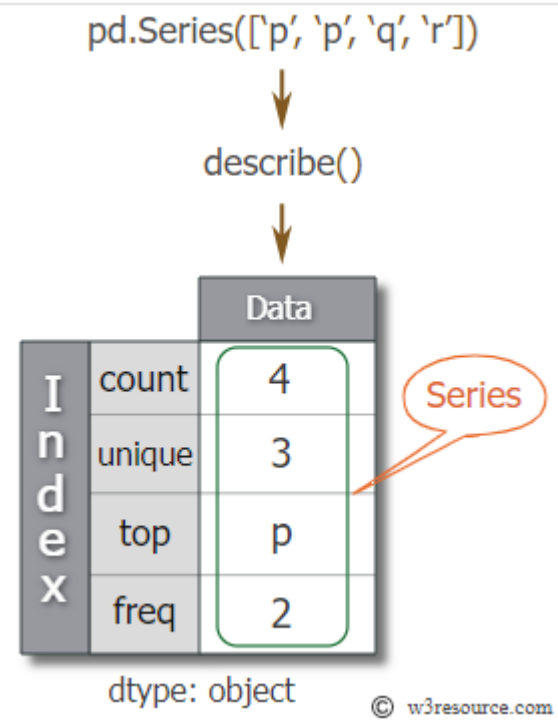


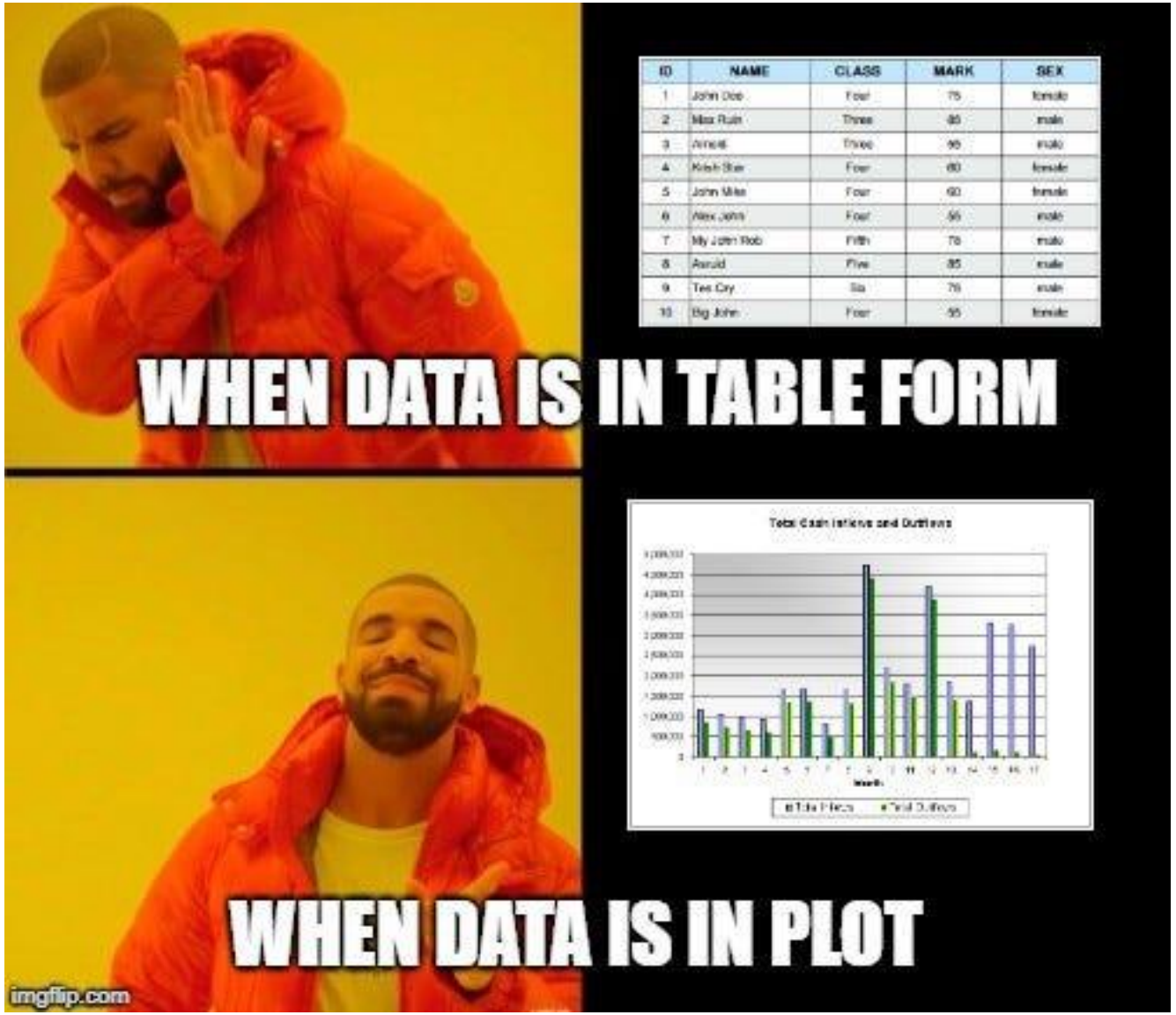
Statystyki opisowe

1. Metody:
   a) describe()
   b) value_counts()
   c) unique() / nunique()



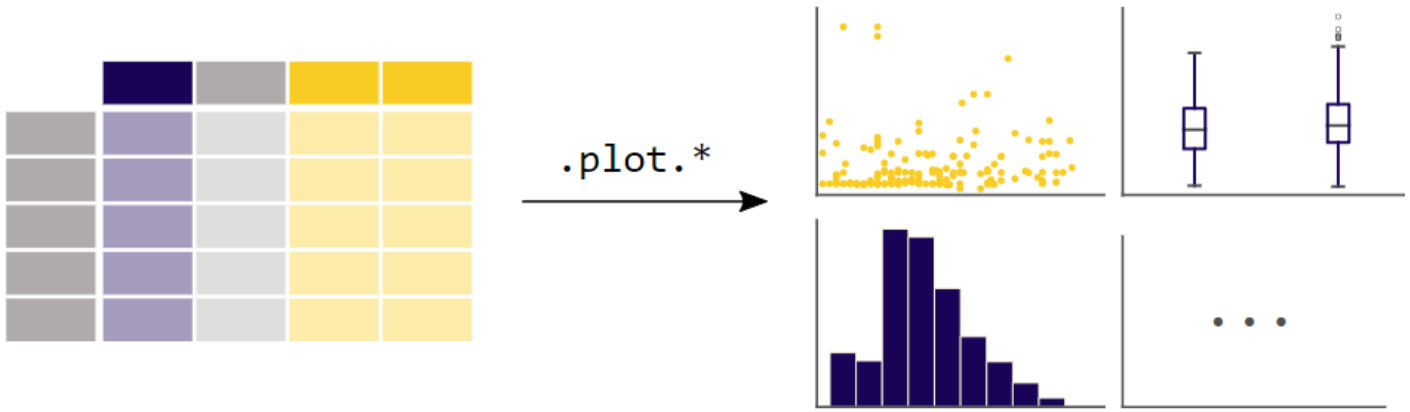


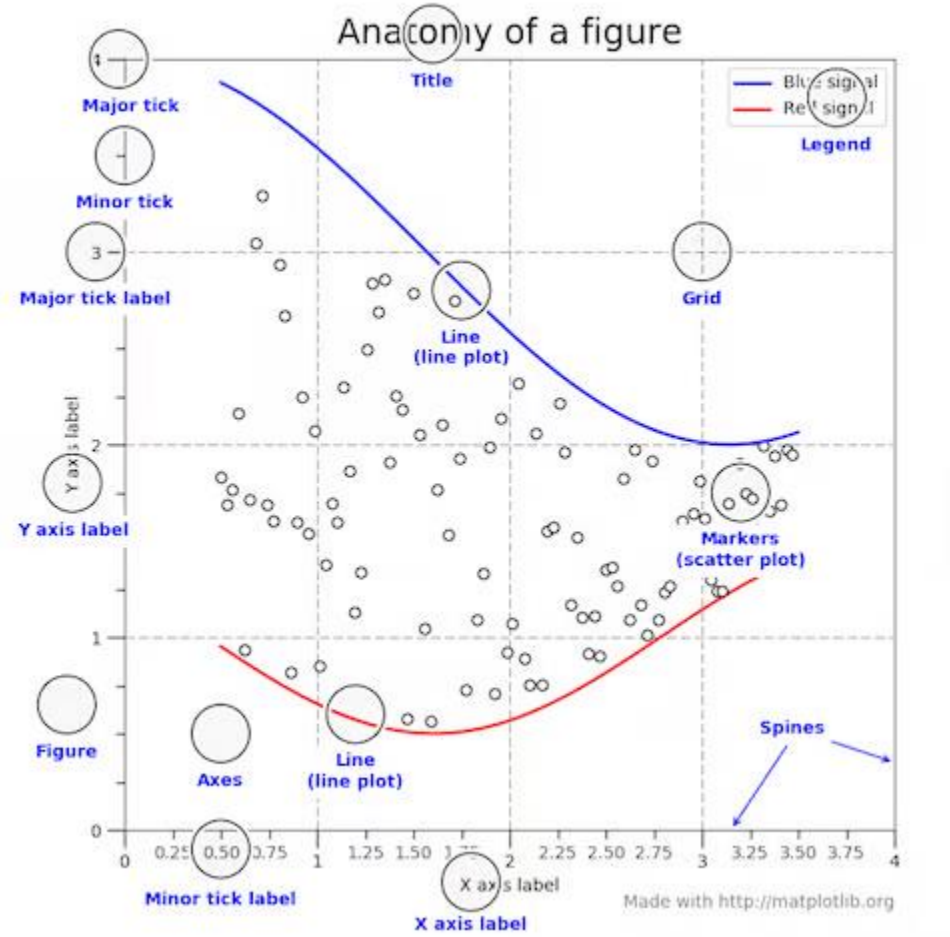



ID NAME CLASS MARK SEX
1 John Doe Four 75 female
2 Max Ruin Three 85 male
3 Arnold Three 90 male
4 Krish Star Four 60 female
5 John Mike Four 60 female
6 Alex John Four 55 male
7 My John Rob Fifth 78 male
8 Asruid Five 85 male
9 Tes Cry Six 75 male
10 Big John Four 55 female
WHEN DATA IS IN TABLE FORM
WHEN DATA IS IN PLOT
imgflip.com

Podstawowe wykresy pandas

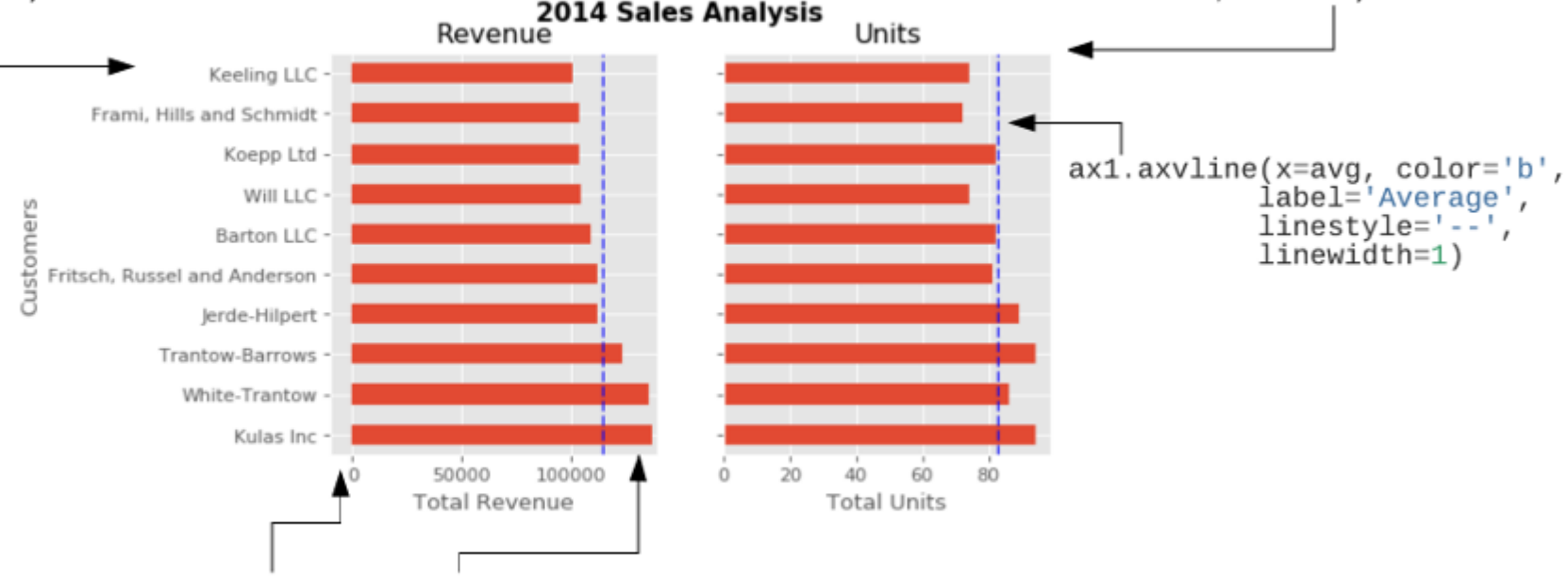
- 'bar' or 'barh' for bar plots
- 'hist' for histogram
- 'box' for boxplot
- 'kde' or 'density' for density plots
- 'area' for area plots
- 'scatter' for scatter plots
- 'hexbin' for hexagonal bin plots
- 'pie' for pie plots





https://pandas.pydata.org/docs/user_guide/visualization.html

```
fig, (ax0, ax1) = plt.subplots(nrows=1,ncols=2, sharey=True, figsize=(7, 4))

fig.suptitle('2014 Sales Analysis', fontsize=14, fontweight='bold')

top_10.plot(kind='barh', y="Sales",
            x="Name", ax=ax0)

top_10.plot(kind='barh', y="Purchases",
            x="Name", ax=ax1)

ax1.axvline(x=avg, color='b',
            label='Average',
            linestyle='--',
            linewidth=1)

ax0.set_xlim([-10000, 140000])

ax0.set(title='Revenue', xlabel='Total
        Revenue', ylabel='Customers')

ax1.set(title='Units',
        xlabel='Total Units', ylabel='')

fig.savefig('sales.png', transparent=False, dpi=80, bbox_inches="tight")
```



pbpython.com

https://pbpython.com/effective-matplotlib.html