Python średnio zaawansowany

Dzień 12

Blok nr 4:

Analiza danych

AGENDA

- Podstawowe metody DataFrame
- Wyszukiwanie wartości minimalnych, maksymalnych
- Średnia, mediana
- Dlaczego średnia kłamie?
- Odchylenie standardowe
- Percentyle
- Rozkład Gaussa, reguła trzech sigm

Dokumentacja

Przykłady oparte są o dokumentację ramek danych modułu pandas:

```
https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html
```

Podstawowe metody DataFrame

Podstawowe metody

`dataframe`

- `.head()` – zwraca pierwsze wiersze (domyślnie: 5)
- `.tail()` – zwraca ostatnie wiersze (domyślnie: 5)
- `.drop(„column_name”)` – usuwa kolumnę
- `.copy()` – zwraca kopię obiektu DataFrame
- `.count()` – zwraca liczbę wierszy

Wartość minimalna, maksymalna

Wartość maksymalna i minimalna

Metoda `max()`:

```
planets_df.rotation_period.max()
```

wyszuka i zwróci maksymalną wartość w obiekcie „**planets_df**”, kolumnie „**rotation_period**”.

Metoda `min()` – przez analogię: zwróci wartość minimalną we wskazanym obiekcie.

Średnia, mediana, odchylenie standardowe

Wartość średnia, mediana

Średnia arytmetyczna:

```
planets_df.rotation_period.mean()
```

Mediana:

```
planets_df['rotation_period'].median()
```

Dlaczego średnia kłamie?

Średnie oceny uczniów z klasy 5a są następujące: 1,1,2,2,3,3,4,4,5,5,6,6

Średnie oceny uczniów z klasy 5b są następujące: 3,3,3,3,3,3,4,4,4,4,4,4

Jaka będzie średnia dla ocen z obu klas?

Dlaczego średnia kłamie?

Jeżeli wyliczylibyśmy średnie ocen dla całej klasy otrzymalibyśmy w obu przypadkach średnią równą 3.5.

Czy takie wartości średnich oddają charakter zebranych ocen?

Źródło: http://www.naukowiec.org/wiedza/statystyka/odchylenie-standardowe_703.html

Odchylenie standardowe

Jest miarą odległości poszczególnych wyników od średniej; czy rozrzut wyników wokół średniej jest niewielki czy wielki?

Teoria:

$$STD = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}}$$

$\bar{X}$ - średnia arytmetyczna

X_i - kolejna obserwacja

N – liczba obserwacji

Praktyka:

```
planets_df.rotation_period.std()
```

Percentyle

Percentyl

Teoria:

https://pl.wikipedia.org/wiki/Percentyl

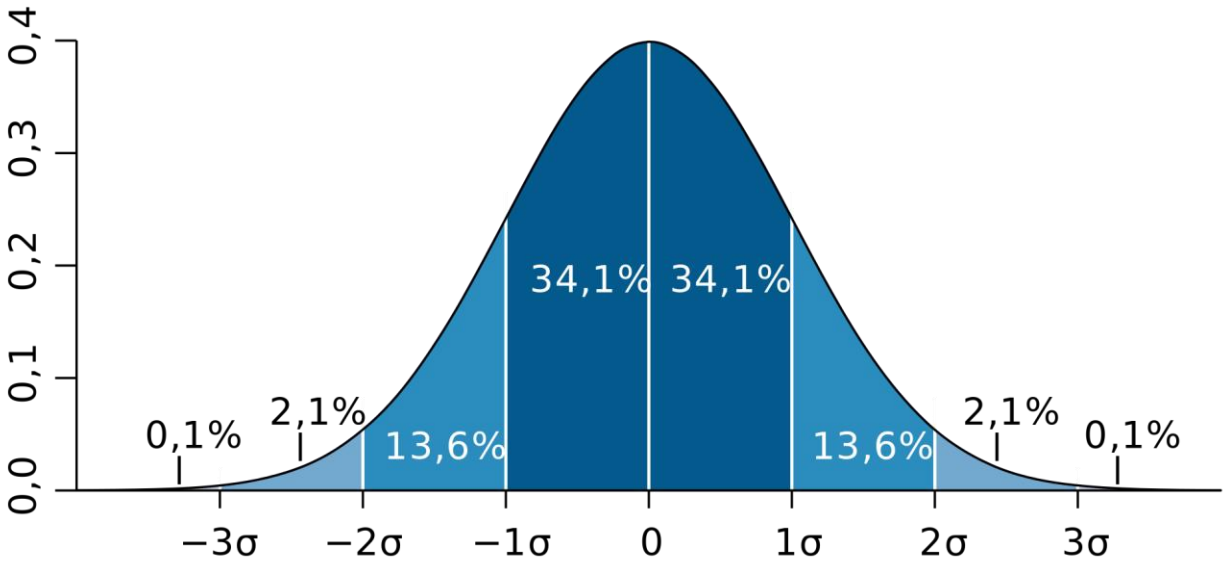
http://www.naukowiec.org/wiedza/statystyka/centyl-percentyl_690.html

Praktyka:

```
planets_df.rotation_period.describe(percentiles=[0.9, 0.95, 0.99])
```

Rozkład Gaussa, reguła trzech sigm

Rozkład Gaussa, reguła trzech sigm



Rozkład Gaussa, reguła trzech sigm

Teoria:

https://pl.wikipedia.org/wiki/Odchylenie_standardowe#Odchylenie_a_obserwacje_dalekie_od_%C5%9Bredniej

http://www.naukowiec.org/wiedza/statystyka/regula-trzech-sigm_709.html

https://www.statystyka-zadania.pl/regula-trzech-sigm/

Rozkład Gaussa, reguła trzech sigm

Praktyka:

```
oferty [ oferty.cena > oferty.cena.mean() + 3 * oferty.cena.std() ]
```

Zwraca oferty spoza 99.7% obserwacji.

Describe()

Zwraca podstawowe statystyki dla ramki danych

```
oferty_df.przebieg.describe()
count    2.548000e+03
mean     1.888843e+05
std      1.065169e+05
min      0.000000e+00
25%      1.400000e+05
(...)

oferty_df.przebieg.describe()['25%’]
140000.0
```

Dzięki!