


Python eksploracja danych

Pracownia programowania 2



Środowisko Jupyter to interaktywne narzędzie, które pozwala na tworzenie i udostępnianie dokumentów, zawierających żywy kod, równania, wizualizacje oraz tekst narracyjny. Jest to otwarte oprogramowanie, szeroko stosowane do analizy danych, uczenia maszynowego, wizualizacji danych, pracy naukowej oraz edukacji. Jupyter wspiera wiele języków programowania, takich jak Python, R, Julia i wiele innych, dzięki czemu jest wszechstronnym narzędziem dla programistów i naukowców z różnych dziedzin.



Notatniki Jupyter (Jupyter Notebooks): To interaktywne dokumenty, które mogą zawierać kod, wykresy, animacje, tekst opisowy, równania matematyczne i inne elementy multimedialne. Notatniki te są bardzo popularne w dziedzinie analizy danych i nauki o danych, ponieważ pozwalają na łatwe eksperymentowanie z kodem i wizualizację danych.

Jupyter Notebook

- Jupyter Notebook jest starszą, bardziej ograniczoną wersją środowiska, która skupia się na pracy z pojedynczymi notatnikami.
- Umożliwia tworzenie i edycję notatników, które zawierają kod, wykresy, tekst narracyjny oraz równania matematyczne.
- Interfejs użytkownika jest stosunkowo prosty i skoncentrowany na edycji i uruchamianiu kodu w ramach jednego notatnika.
- Notebooki są uruchamiane w nowych kartach przeglądarki jako oddzielne instancje.

Jupyter Lab

- JupyterLab to bardziej zaawansowane i elastyczne środowisko, które można postrzegać jako następcę Jupyter Notebook. Zostało zaprojektowane, aby zapewnić zintegrowane środowisko pracy z notatnikami, kodem, danymi i plikami.
- JupyterLab oferuje modułowy interfejs użytkownika, który umożliwia jednoczesne otwieranie i edycję wielu notatników oraz plików w jednym oknie przeglądarki, dzięki czemu praca jest bardziej zorganizowana i efektywna.
- Wprowadza koncepcję obszarów roboczych
- Oferuje wbudowaną obsługę przeglądarki plików, panelu wizualizacji, edytora tekstu oraz terminala, co czyni go bardziej kompleksowym narzędziem.
- JupyterLab jest rozszerzalny, co oznacza, że użytkownicy mogą instalować dodatki, które dodają nowe funkcjonalności lub integrują się z zewnętrznymi usługami i narzędziami.

DataFrame to dwuwymiarowa struktura danych podobna do tabeli w bazie danych lub arkusza kalkulacyjnego Excela. DataFrame składa się z wierszy i kolumn – każda kolumna w DataFrame to Series.

Wybieranie kolumn: **df['nazwa_kolumny']**

Filtrowanie: **df[df['nazwa_kolumny'] > 60]**

Sortowanie: **df.sort_values('count')**

Dodawanie nowej kolumny:

```
df['amount'] = [1200, 3532, 24000, 2000]
```

Usuwanie kolumny:

```
df = df.drop('amount', axis=1)
```

Zmiana nazwy kolumny:

```
df = df.rename(columns={'deviceId': 'id', 'count': 'quantity'})
```

Funkcje **apply** i **map** pozwalają na zastosowanie wybranej funkcji do każdego elementu zapisanego w Series lub DataFrame.

```
df['Fare_test'] = df['Fare'].apply(np.sqrt)
```

Funkcja **fillna()** wypełnienie brakujących danych określoną wartością

```
df = df.fillna(value = 0)
```

```
df = df.fillna(df['Age'].median())
```

Usunięcie wybranych wierszy **dropna()**

```
df = df.dropna()
```

Usunięcie duplikatów **drop_duplicates()**

```
df = df.drop_duplicates()
```

Zmiana typu **astype()**

```
df['count'] = df['count'].astype(int)
```



Funkcja **groupby()** do pogrupowania wartości po kolumnie

```
grouped = df.groupby('product_id')
```

Wczytywanie i zapisywanie danych:

`read_csv()`, `to_csv()`, `read_excel()`, `to_excel()`, `read_sql()`, `to_sql()`

Wybieranie danych:

`.loc[]`, `.iloc[]`

Manipulacja danymi:

`drop()`, `rename()`, `set_index()`, `reset_index()`, `pivot()`, `melt()`

Czyszczenie danych:

`dropna()`, `fillna()`, `replace()`, `drop_duplicates()`

Analiza danych:

`describe()`, `value_counts()`, `groupby()`, `corr()`

Statystyki:

`mean()`, `median()`, `min()`, `max()`, `std()`, `quantile()`

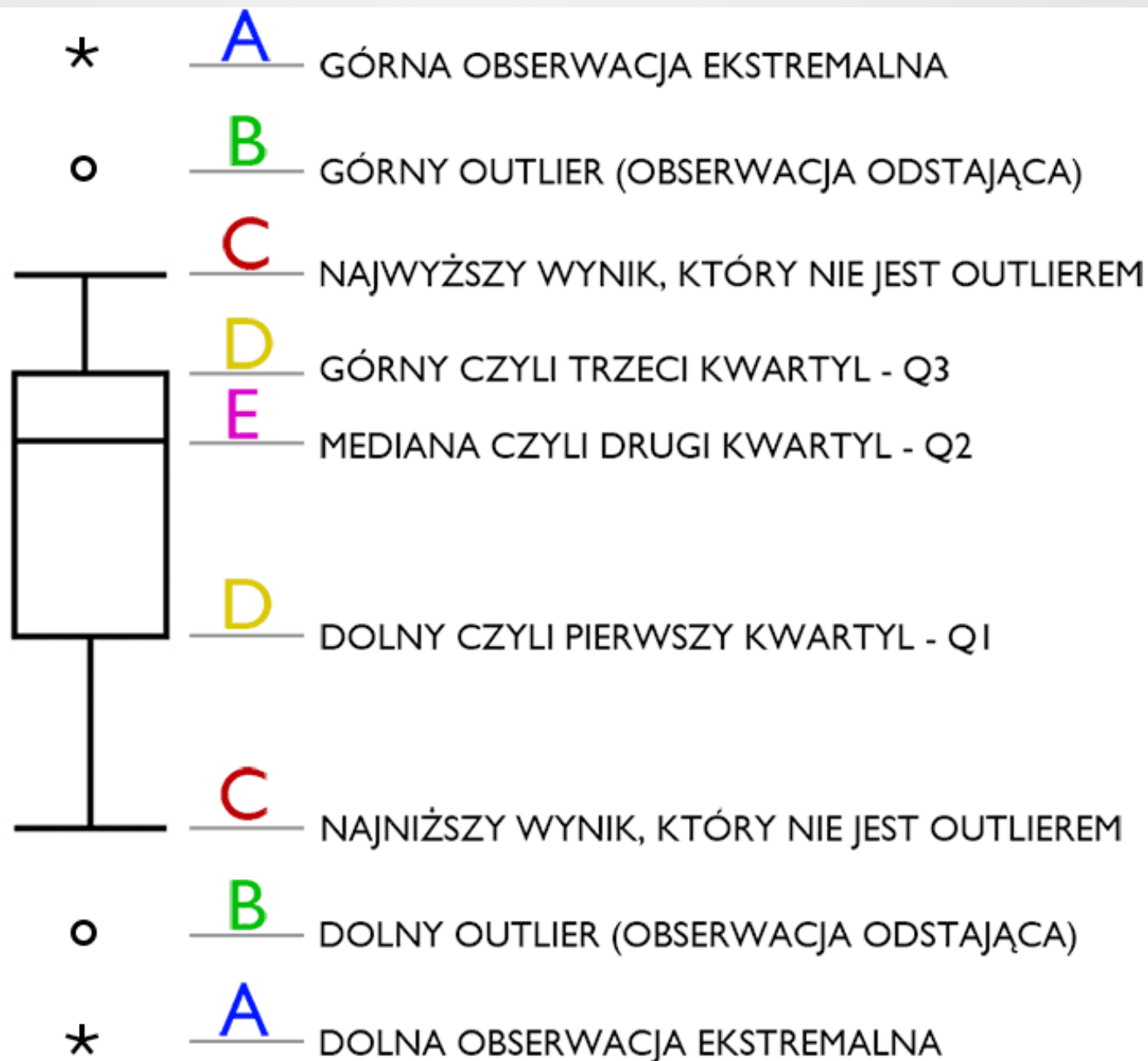
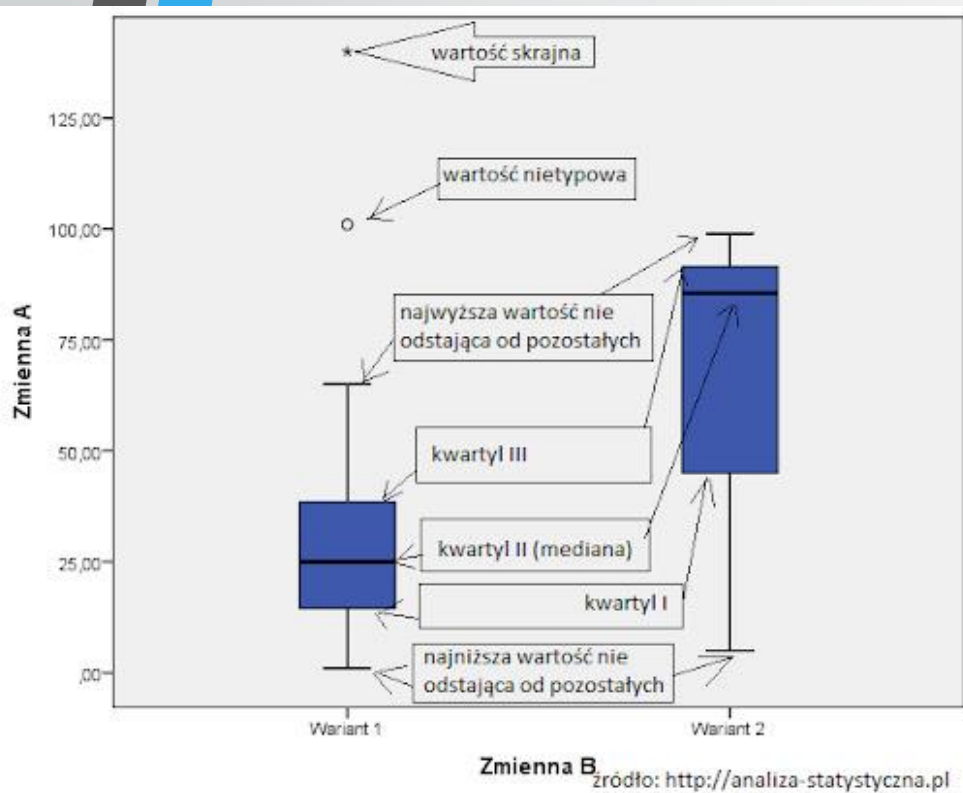
Operacje na ciągach znakowych:

`str.lower()`, `str.upper()`, `str.contains()`, `str.replace()`, `str.split()`, `str.join()`

Wizualizacja danych:

`plot()`, `hist()`, `boxplot()`

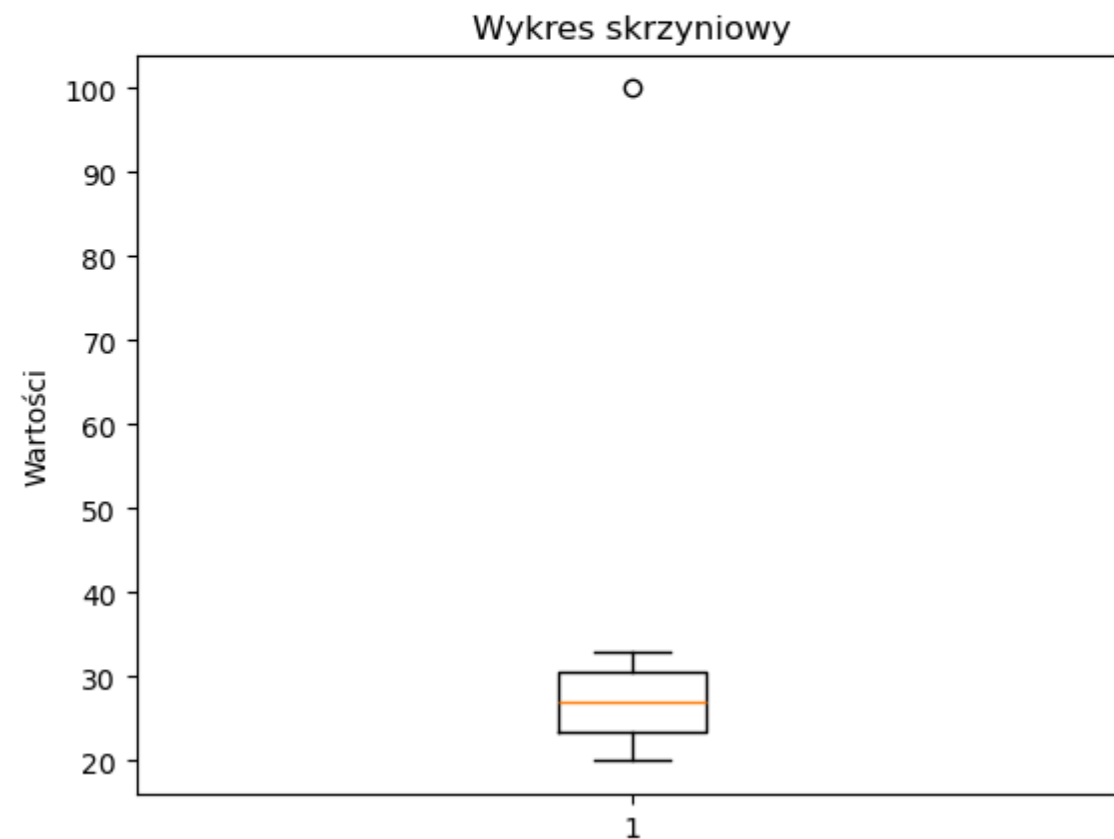
Wykres skrzyniowy



```
import matplotlib.pyplot as plt

# Przykładowe dane
data = [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 100]

# Tworzenie wykresu skrzyniowego
plt.boxplot(data)
plt.title('Wykres skrzyniowy')
plt.ylabel('Wartości')
plt.show()
```



Zadanie 1: Wstępne przetwarzanie danych

1. Wczytaj zbiór danych titanic.csv do DataFrame'u przy użyciu biblioteki pandas.
2. Wyświetl pierwsze 10 wierszy zbioru danych.
3. Sprawdź, czy w danych występują brakujące wartości. Jakie kolumny zawierają brakujące dane?
4. Wypełnij brakujące wartości w kolumnie Age medianą wieku. Dla kolumny Embarked użyj najczęściej występującej wartości.

Zadanie 2: Eksploracja danych

1. Ile było pasażerów na pokładzie Titanica (w zbiorze danych)?
2. Ile było kobiet i ile mężczyzn na pokładzie?
3. Jakie było średnie przeżycie wśród kobiet a jakie wśród mężczyzn?
4. Stwórz nową kolumnę IsChild, która będzie miała wartość True jeśli pasażer miał mniej niż 18 lat, a w przeciwnym wypadku False. Ile było dzieci na pokładzie?

Zadanie 3: Wizualizacja danych


1. Utwórz wykres pokazujący rozkład wieku pasażerów. Użyj histogramu.
2. Zbuduj wykres przedstawiający stosunek liczby przeżyć do liczby zgonów.
3. Stwórz wykres pokazujący zależność między klasą biletu (Pclass) a przeżywalnością.
4. Wykonaj wykres skrzyniowy (boxplot) prezentujący rozkład cen biletów (Fare) w zależności od klasy biletu (Pclass).

Zmienna fikcyjna (ang. dummy variable)

Zmienna fikcyjna nazywana jest również zmienną binarną, ślepą, formalną lub zerojedynekową. Zmienna ta mierzona jest na skali nominalnej lub porządkowej, której można przypisać wyłącznie wartości binarne (zerojedynekowe), oznaczające, że zmienna albo posiada jakąś wartość albo jej nie posiada. Przykładowe wartości takiej zmiennej mogą wyglądać następująco:

- 1.dla zmiennej „płeć”: kobieta – mężczyzna
- 2.dla zmiennej „status zawodowy”: pracujący – niepracujący
- 3.dla zmiennej „miejsce zamieszkania”: mieszkaniec miasta – mieszkaniec wsi.

Dummy coding – w statystyce jest to metoda kodowania danych nominalnych (jakościowych) na dane liczbowe przyjmujące wartości 0 i 1 w celu ich analizy statystycznej.



W praktyce, proces budowy modelu predykcyjnego często zaczyna się od zdefiniowania, jakie są cechy (wejścia) i co jest etykietą (wyjście), ponieważ jest to kluczowe dla ustalenia, jakiego typu uczenia maszynowego użyć i jak przygotować dane do treningu. Cechy są używane do "nauczenia" modelu wzorców, które pozwalają na dokonanie predykcji lub klasyfikacji, natomiast etykieta jest używana w procesie treningu do oceny, jak dobrze model radzi sobie z przewidywaniami i do dostosowania jego parametrów.



Cechy (Features)

Cechy to indywidualne niezależne zmienne, które są wejściem do modelu. W kontekście zbioru danych, każda kolumna reprezentująca różne atrybuty lub właściwości obserwacji może być traktowana jako cecha. Na przykład, w zbiorze danych dotyczących domów, cechy mogą obejmować powierzchnię domu, liczbę sypialni, wiek budynku, odległość od centrum miasta itd. W kontekście analizy statystycznej, cechy są zmiennymi, na podstawie których przewiduje się wartości innych zmiennych lub analizuje się zależności między zmiennymi.



Etykieta (Label)

Etykieta to zmienna, którą model próbuje przewidzieć. W uczeniu nadzorowanym, każda obserwacja w zbiorze danych zawiera etykietę, która jest prawdziwą wartością wyjściową dla danej obserwacji. Etykieta jest zależną zmienną, która może przyjmować różne formy, w zależności od problemu, jaki model ma rozwiązać:

The image features a central graphic of four concentric circles in varying shades of gray, creating a tunnel-like effect. In the center of these circles, the words "The End" are written in a white, elegant, cursive script. To the left of the circles, there is a blue geometric shape that resembles a stylized corner or a folded piece of paper, with a white outline and a blue fill. The overall composition is clean and modern, with a focus on geometric shapes and a classic script font.

The End