

Eksploracyjna analiza danych

Paweł Gliwny

Podstawowe statystyki

- **Statystyki zbiorcze** pozwalają nam zrozumieć właściwości pewnego zbioru danych.
 - Dane nie zawsze mają postać zbioru danych albo arkusza kalkulacyjnego.
 - Często pojawiają się w postaci statystyk zbiorczych.
- **Statystyka opisowa** pierwszy krok do zrozumienia, co dane pokazują i gdzie może być potrzebna głębsza analiza.
 - Dostarcza podsumowania zbioru danych za pomocą **metryk** takich jak średnia, mediana i odchylenie standardowe.

Średnia, mediana i moda

- **Średnia (arytmetyczna)** to suma wszystkich wartości podzielona przez liczbę tych wartości.
 - Pozwala stwierdzić, ile każda obserwacja w serii wносиłaby do ogólnej sumy, gdyby każda obserwacja generowała jednakową wielkość.
- **Mediana** to środkowy punkt całego zakresu danych, gdybyśmy posortowali je w kolejności.
- **Moda** to liczba występująca najczęściej w zbiorze danych.

Miary

- Średnia, mediana i moda nazywane są **miarami lokalizacji** lub **miarami tendencji centralnej**.
- **Miary zmienności** — wariancja, rozpiętość i odchylenie standardowe — to miary rozrzutu.
- Lokalizacja mówi, w którym miejscu osi liczbowej wypada typowa wartość, a rozrzut — jak daleko znajdują się inne liczby od tej wartości.

Przykład

- Nasz dane: [7, 5, 4, 8, 4, 2, 9, 4, 100]
- Średnia: 15,89
- Mediana: 5
- Moda: 4

*Średnia 15,89 jest liczbą, która **nie pojawia się w danych**.*

Koszykarz LeBron James zdobywa średnio 27,1 punktu na mecz.

Częste pomyłki

~~Średnia reprezentuje środkowy punkt danych (którym jest mediana).~~

Połowa liczb znajduje się powyżej średniej, a połowa poniżej.

- **To nieprawda.**
- Często zdarza się, że większość danych znajduje się poniżej (lub powyżej) średniej.
 - Na przykład ogromna większość ludzi ma liczbę palców większą od średniej (która zapewne wynosi 9 z kawałkiem).

Projekty związane z danymi

- Nigdy nie są proste
- Interesariusze zwykle widzą prezentację *Power Point*
 - Według sztywnego skryptu: od pytania, poprzez dane do odpowiedzi.

Czego nie widać?

- Wszystkich idei które nie przeszły przez sito - **ważnych decyzji i założeń**, które podjął zespół aby uzyskać odpowiedź.

Eksploracyjna analiza danych

Dobry zespół ds. danych podąża nie prostą, ale krętą ścieżką, dostosowując się do dokonywanych odkryć. W miarę podróży wraca do wcześniejszych pomysłów i zauważa, że w rezultacie otworzyło się wiele nowych dróg.

- Ten proces iteracji, odkryć i przyglądania się danym nosi nazwę **eksploracyjnej analizy danych** (ang. exploratory data analysis, eda)

Eksploracyjna analiza danych

- Sformułowany przez statystyka **Johna Tukeya** w latach 70.
- Sposób na wstępne zrozumienie danych poprzez **zbiorcze statystyki i wizualizacje** przed zastosowaniem bardziej skomplikowanych metod.
- Tukey postrzegał EDA jako pracę detektywistyczną.
- W danych ukryte są wskazówki, a właściwa eksploracja może zasugerować następne kroki.

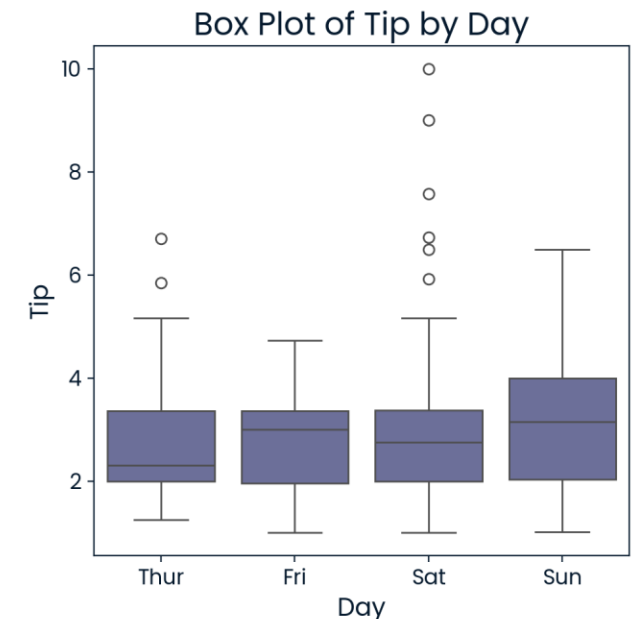
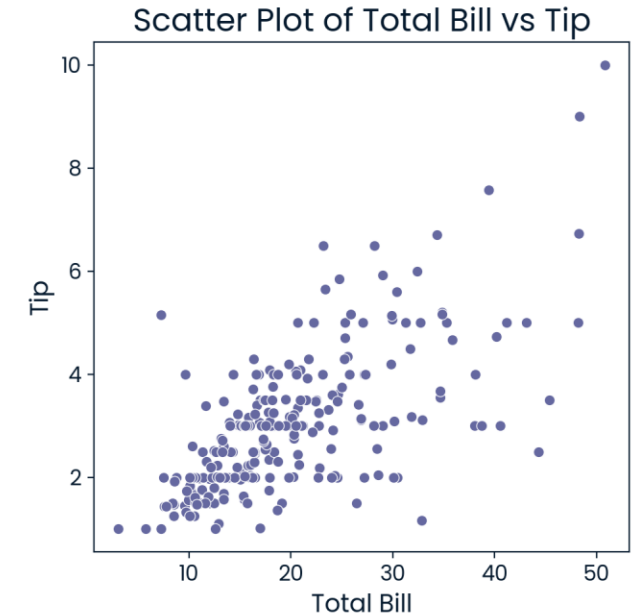


Badanie Eksploracyjne

- Względnie **nowe zagadnienie** w statystyce
- **Klasyczna statystyka** skupia się na wnioskach, czasami przy użyciu skomplikowanych procedur aby **wyciągnąć wnioski o dużych populacjach na podstawie niewielkich prób**.
- 1962 John W. Turkey w artykule *The future of Data Analysis* wezwał do reformy statystyki.
- Zaproponował nową dyscyplinę o nazwie **analiza danych**, której **tylko jednym z elementów** byłoby wnioskowanie statystyczne.

Analiza danych według Turkeya

- Zakres badań eksploracyjnych został przedstawiony w książce ***Exploratory Data Analysis*** [Turkey 1977]
- Zaprezentowano tam proste wykresy (np. pudełkowy czy punktowy)
- Wykresy wraz ze **statystykami podsumowującymi** (średnia, mediana, kwantyle) pomagają zobrazować własności danych.
- Oryginalne tezy Turkeya są zaskakująco trwałe i tworzą częściową podstawę *data science*.



Klasyczne podejście vs EDA

Klasyczne podejście:

- Statystyka = głównie wnioskowanie (testy hipotez, przedziały ufności)
- Cel: wyciągnąć formalne wnioski o populacji

Wizja Tukeya (EDA):

Analiza danych = szersza dziedzina obejmująca:

- **Eksplorację** (oglądanie, rozumienie danych)
- **Wizualizację**
- **Wykrywanie wzorców**
- **Czyszczenie danych**
- **Wnioskowanie statystyczne** (jako jeden z elementów, nie jedyny cel)

Percentyle

- **Percentyl P** to wartość, dla której:
 - Co najmniej **$P\%$** obserwacji ma wartość **mniejszą lub równą**
 - Co najmniej **$(100-P)\%$** obserwacji ma wartość **większą lub równą**

Jak obliczyć?

- **Posortuj dane** od najmniejszej do największej wartości
- **Znajdź pozycję** odpowiadającą $P\%$ długości zbioru
- **Odczytaj wartość** na tej pozycji

Przykład: 80. percentyl

Dla zbioru $\{3, 1, 5, 3, 6, 7, 2, 9\}$:

- Po sortowaniu: $\{1, 2, 3, 3, 5, 6, 7, 9\}$
- 80% drogi od początku \rightarrow wartość między 7 a 9

Kwantyle, Kwartyle i Percentyle

- **Kwantyl** to ogólna nazwa dla wartości, która dzieli posortowany zbiór danych na części w określonej proporcji.
- Kwantyl zapisujemy jako **ułamek lub liczbę z przedziału $[0,1]$**
- Przykłady:
 - Kwantyl **0.25** = 25. percentyl = Q1
 - Kwantyl **0.50** = 50. percentyl = mediana = Q2
 - Kwantyl 0.75 = Q3
 - Kwantyl **0.80** = 80. percentyl
- **Kwartyl** - specjalny przypadek (tylko Q1, Q2, Q3)

Wykres pudełkowy

- **Mediana** = 50. percentyl (środek zbioru)
- **Q1** = 25. percentyl (pierwszy kwartył)
- **Q3** = 75. percentyl (trzeci kwartył)
- Rozstęp międzykwartyłowy (**IQR**)

$$\text{IQR} = Q3 - Q1$$

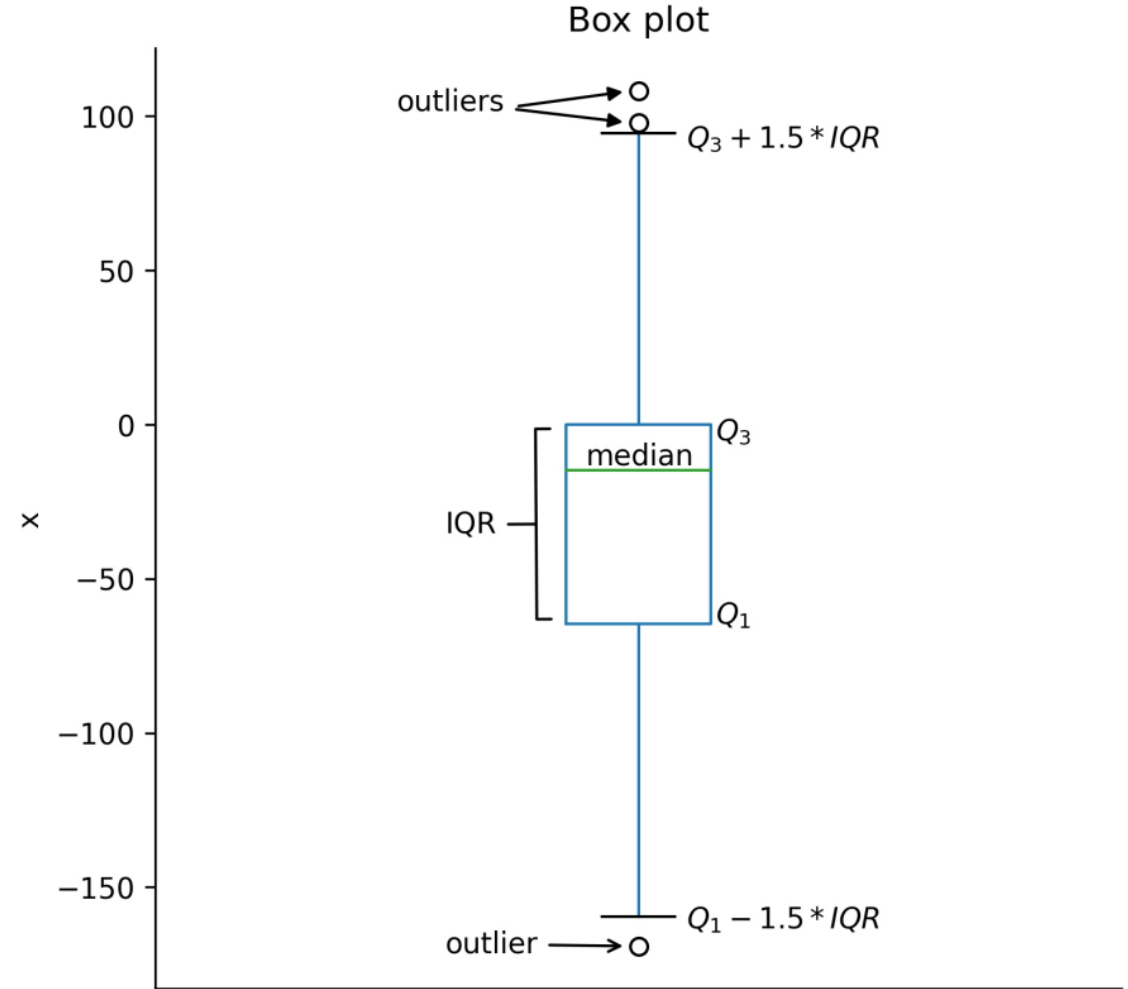


Figure 1.6 – The Tukey box plot

Wizualizacja danych

- **Pierwszy etap** układu analizy danych, ułatwiający pojmowanie i przekazywanie informacji.
- Przedstawiamy dane i informacje w formie graficznej przy użyciu **wykresów, schematów i map**.
- Pozwala wychwytywać wzory, trendy, elementy odstające, rozkłady i wzajemne relacje.
- Pozwala w skuteczny sposób radzić sobie z **dużą liczbą danych**.

Matplotlib

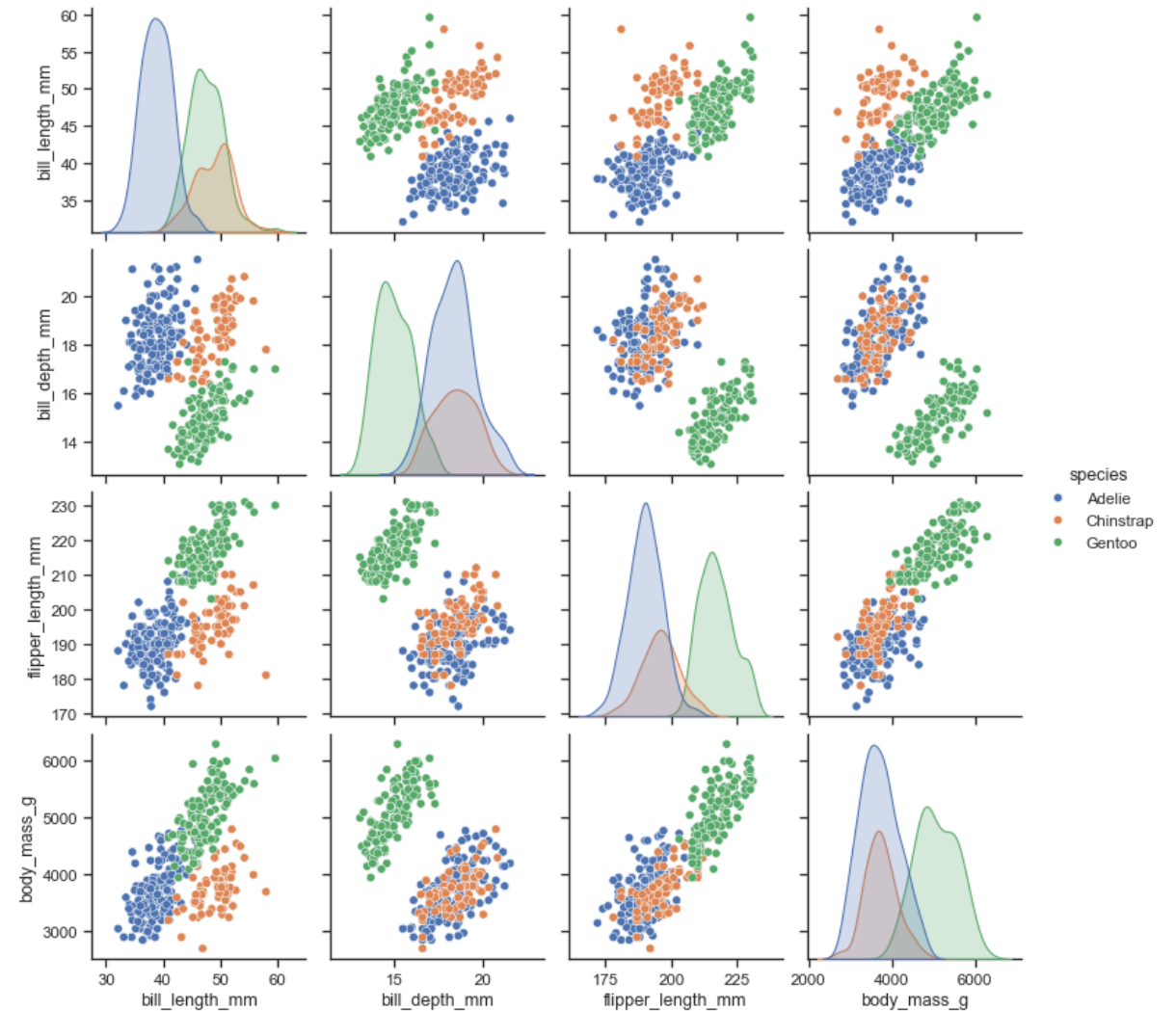
Wszechstronna biblioteka do tworzenia statycznych, animowanych i interaktywnych **wizualizacji w Pythonie**.

Dzięki Matplotlib proste rzeczy są proste, a trudne stają się możliwe.

- Tworzenie **wykresów o jakości publikacyjnej**.
- Dostosowywanie **stylu wizualnego i układu**.
- Osadzanie w **JupyterLab** i **graficznych interfejsach użytkownika**.
- Korzystanie z bogatego zestawu **pakietów zewnętrznych** opartych na Matplotlib.

Seaborn

- Biblioteka do wizualizacji danych w języku Python oparta na matplotlib.
- Zapewnia ona zaawansowany interfejs do rysowania atrakcyjnych i bogatych w informacje wykresów statystycznych.
- Źródło: seaborn.pydata.org



Zadania

- Do zrobienia notebook

Przyjmij nastawienie eksploracyjne

- Dziesiątki dostępnych narzędzi i języków programowania mogą pomóc szybko i niedrogo eksplorować dane poprzez zbiorcze statystyki i wizualizacje.
- **Nie należy jednak myśleć o EDA jak o przyborniku z narzędziami albo liście kontrolnej.**
- Jest to raczej **nastawienie mentalne** wplecione w każdą fazę pracy z danymi, które możesz przyjąć nawet bez zaplecza analitycznego.

Pytania naprowadzające

Choć nie ma jednej właściwej ścieżki, którą należy podążać, jest **kilka pytań**, które możesz zadać, aby pomóc zespołowi w dojściu do użytecznych wniosków:

- Czy dane mogą odpowiedzieć na pytanie?
- Czy odkryliśmy jakieś związki?
- Czy znaleźliśmy w danych nowe możliwości?

EDA I Ty

- EDA może być dla niektórych niekomfortowa — ujawnia ona subiektywną naturę (sztukę?) pracy z danymi.
- Dwa zespoły, otrzymawszy ten sam problem i dane, mogą wybrać **dwie różne ścieżki** analizy
- Czasem dochodząc do tych samych wniosków. A czasem **nie**.

Dlaczego?

EDA I Ty (odpowiedź)

- Po drodze jest do podjęcia zbyt wiele decyzji, żeby dwa zespoły (lub dwie osoby) zrobiły wszystko tak samo.
- Każda osoba będzie miała inne kompetencje, pomysły i narzędzia potrzebne do rozwiązania problemu.

EDA jako ciągły proces, który należy do obowiązków każdego specy od danych, bez względu na to, czy bezpośrednio pracuje z danymi, czy jest członkiem zarządu firmy.

Zapytania do bazy danych vs eksploracja danych

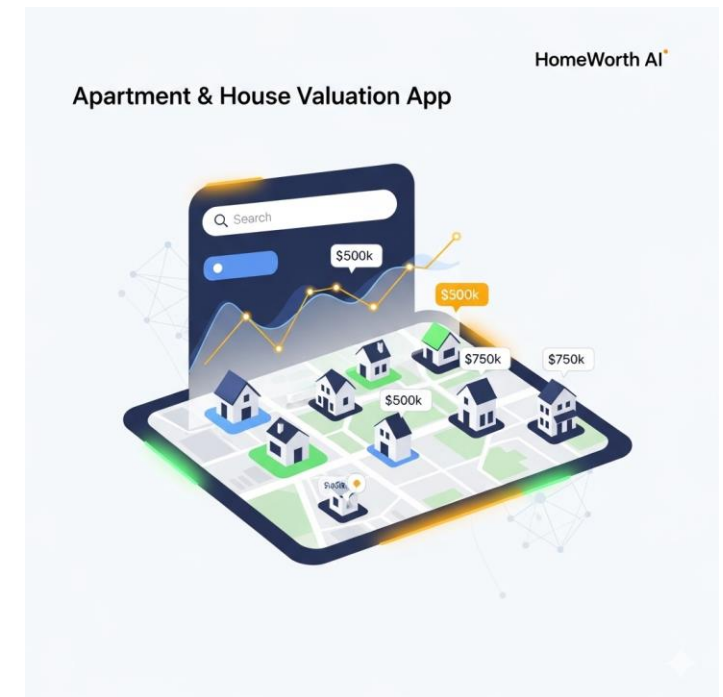
- Zapytania do bazy danych dają odpowiedzi na **konkretne pytania** i zwracają **liczbowe wyniki**.
- Eksploracja danych (ED) to proces **odkrywania wzorców** i zależności w dużych zestawach danych, prowadzący do **nowych wniosków**.
- Zapytania są precyzyjne, ED prowadzi do **głębszej analizy** i odkrywania nieoczywistych informacji.

Zapytania do baz danych vs ED

Zapytanie do bazy danych	Eksploracja danych (ED)
Ile sprzedano coli i wody mineralnej w poszczególne dni tygodnia?	Co kupowali klienci, którzy kupili colę?
Jakie są najczęściej wybierane wycieczki?	Jakie strony odwiedziły osoby po obejrzeniu stron biura podróży?
Jaką najwyższą temperaturę miał pacjent z gripą?	Jakie są objawy grypy?
Jakie były wyniki testów egzaminu?	Jakie cechy uczniów wpływają na wyniki egzaminów?
Ile jest zaobserwowanych asteroid?	Kiedy uderzy w nas asteroida?

Historia – start-up na rynku nieruchomości

- Pracujesz w start-upie działającym na rynku nieruchomości.
- Twoim zadaniem jest **zwiększenie ruchu na stronie**.
- Konkurencja: amerykański gigant **Zillow.com** i jego narzędzie **Zestimate**® (wycena nieruchomości online).
- W Europie podobne funkcje pełnią m.in.:
 - **Otodom, OLX Nieruchomości** (Polska)
 - **Idealista** (Hiszpania, Portugalia, Włochy)
 - **Immoweb** (Belgia)



Przykład wycena domów

- Amerykański serwis Zillow
- Polskie odpowiedniki SonarHome,
- Zbiór danych utworzonym do celów edukacyjnych: Ames Housing Data [kaggle house data](#)

Start-up: twoje zadanie

- Firma potrzebuje **własnego narzędzia predykcyjnego** do wyceny nieruchomości.
- Szef daje Ci zbiór danych:
 - 80 kolumn opisujących cechy domów setki transakcji z lat 2006–2011 (Ames, Iowa).

Cel: przewidywanie ceny sprzedaży na podstawie cech domu.

Pierwsze kroki – zdrowy rozsądek

- Jakie dane powinny wpływać na cenę domu?
 - metraż, liczba pokoi, liczba łazienek, rok budowy, lokalizacja
- Sprawdź, czy takie informacje są w zbiorze danych.
- Oceń, czy dane są wystarczające do zbudowania **sensownego modelu**.

Typy danych w zbiorze

- **Liczbowe:** powierzchnia, rok budowy, liczba pokoi
- **Porządkowe:** ogólna jakość domu (skala 1–10)
- **Kategoryczne:** dzielnica, typ nieruchomości

Już na tym etapie widać, że dane mają potencjał do budowy modelu.

Zakres danych – pułapki

- Sprawdź, **co obejmuje zbiór danych**:
 - Tylko domy jednorodzinne?
 - Brak mieszkań, apartamentów czy bliźniaków?
- Jeśli zakres jest wąski → model będzie miał **ograniczone zastosowanie**.
- W Polsce/EU: duży udział mieszkań w blokach → musisz zadbać o ich uwzględnienie.

Czy wartości mają sens?

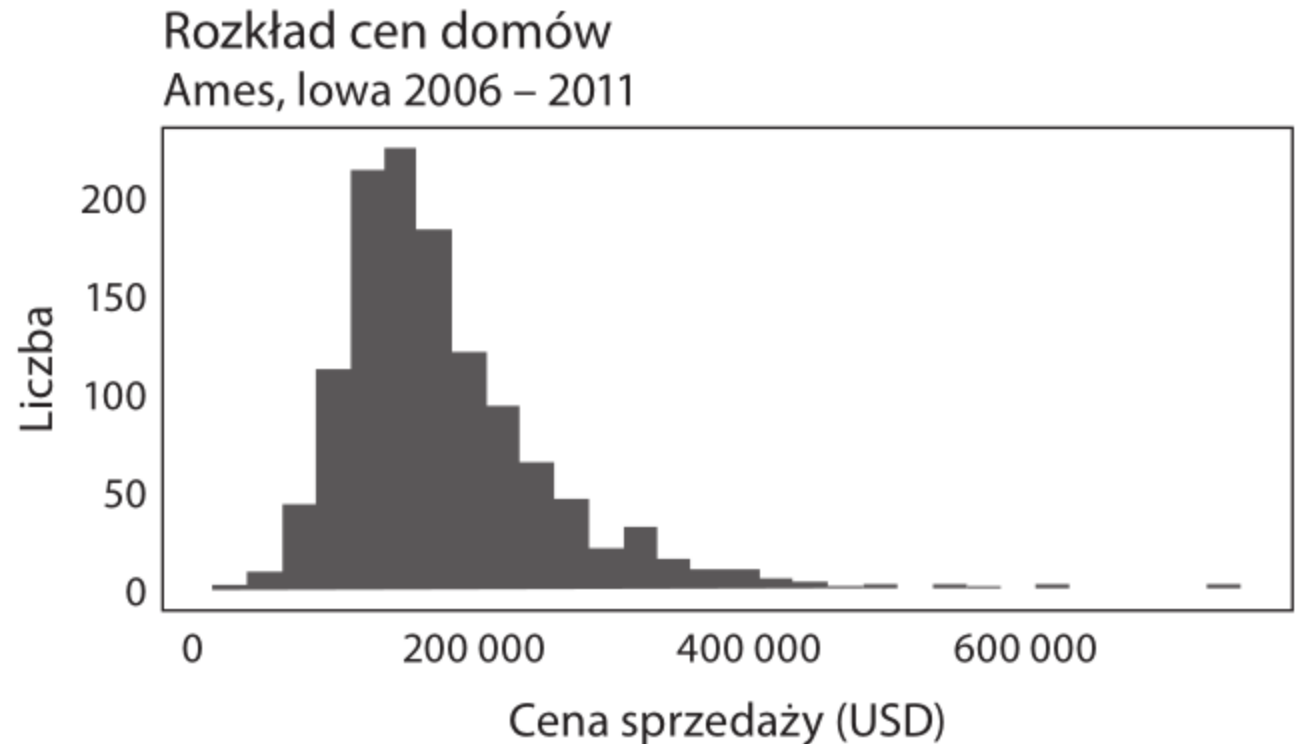
- EDA:
 - statystyki opisowe (średnia, mediana, min/max)
 - wizualizacje (histogramy, boxploty, mapy ciepła)
- Twoim zadaniem jest sprawdzić, czy wyniki są **intuicyjne**.
 - np. czy większy metraż zwykle = wyższa cena?
- Szukaj anomalii i błędów (np. dom o powierzchni 10 m² za milion zł).

Lekcja

- Nie zaczynaj od „magicznych algorytmów”.
- Najpierw sprawdź, czy dane są:
 - sensowne,
 - kompletne,
 - przydatne w kontekście biznesowym.
- Solidna **EDA** = fundament dobrego modelu predykcyjnego.

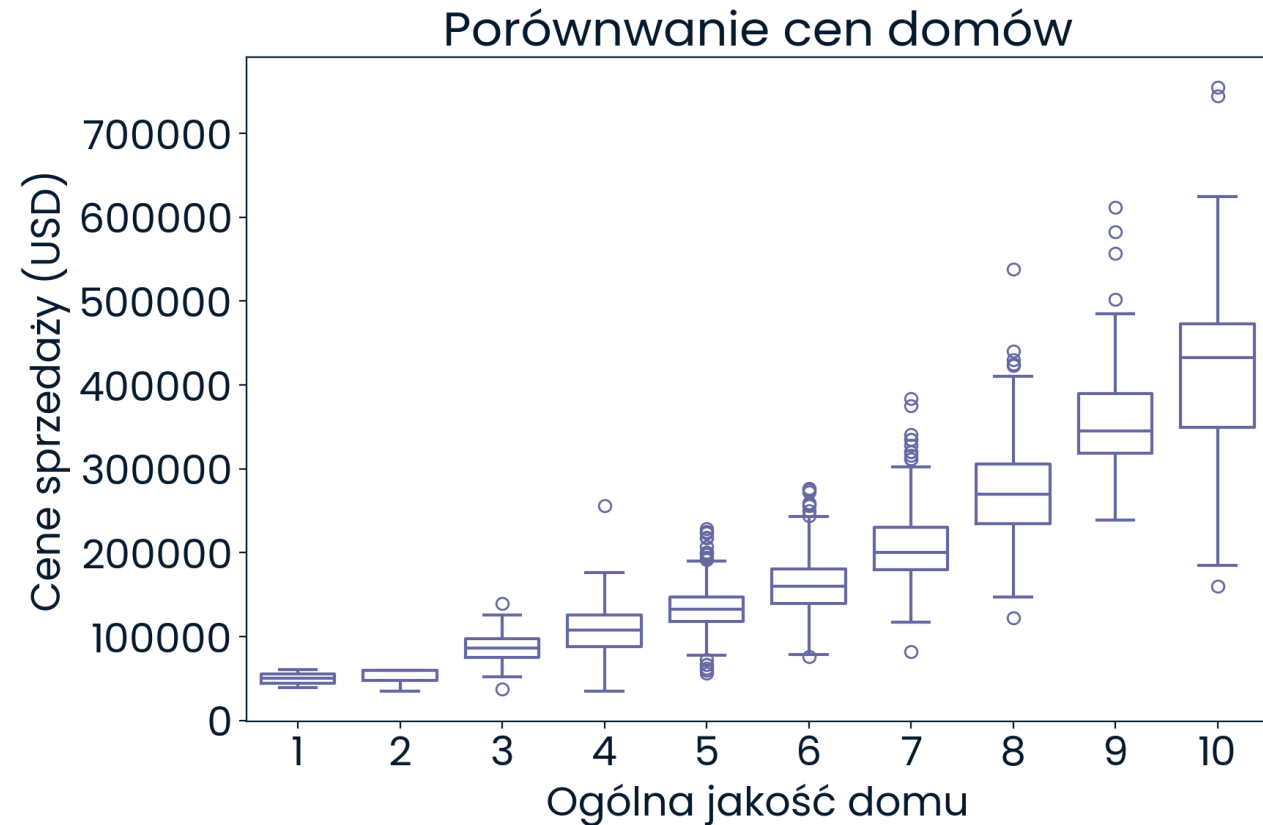
Rozkład cen domów

- Możesz poznać kształt albo rozkład ciągłych danych liczbowych poprzez przyjrzenie się **histogramowi**.
- Histogramy pomagają dostrzegać anomalie.



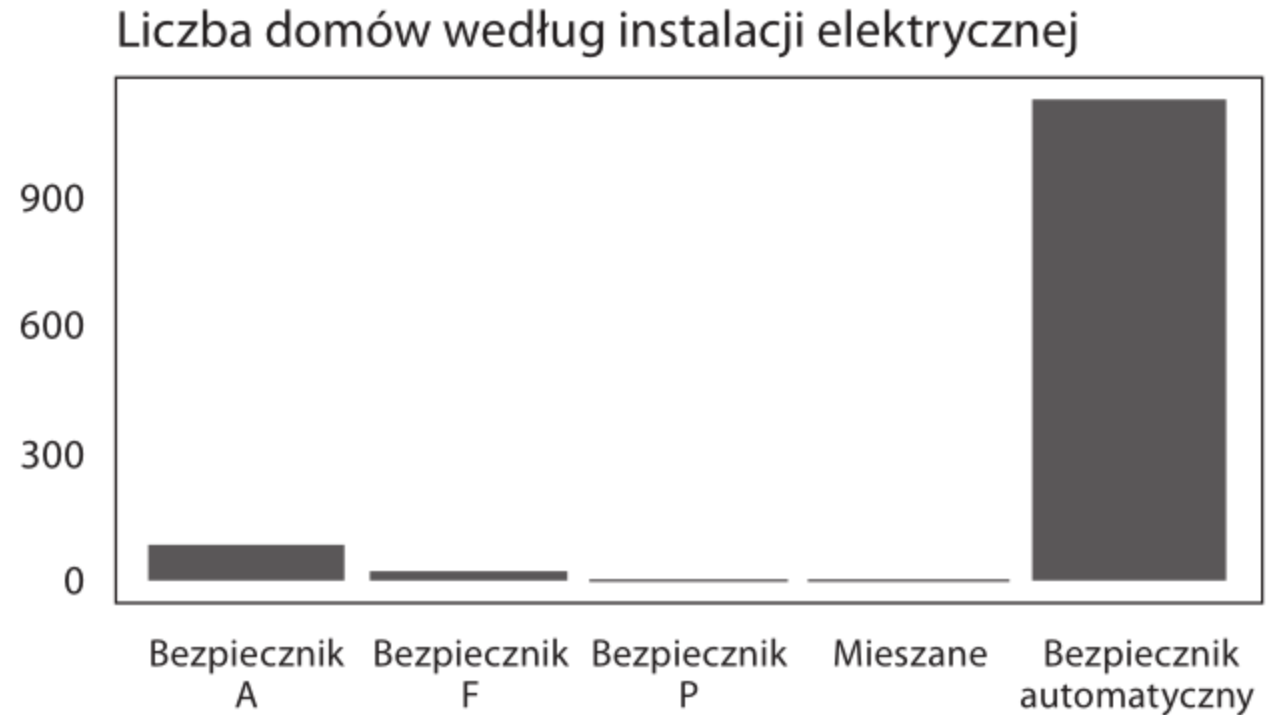
Wykres skrzynkowy

- Można wykorzystać do porównywania danych w kilku grupach.
- Dom z "10" poniżej 200 tys \$, pewnie inne powody



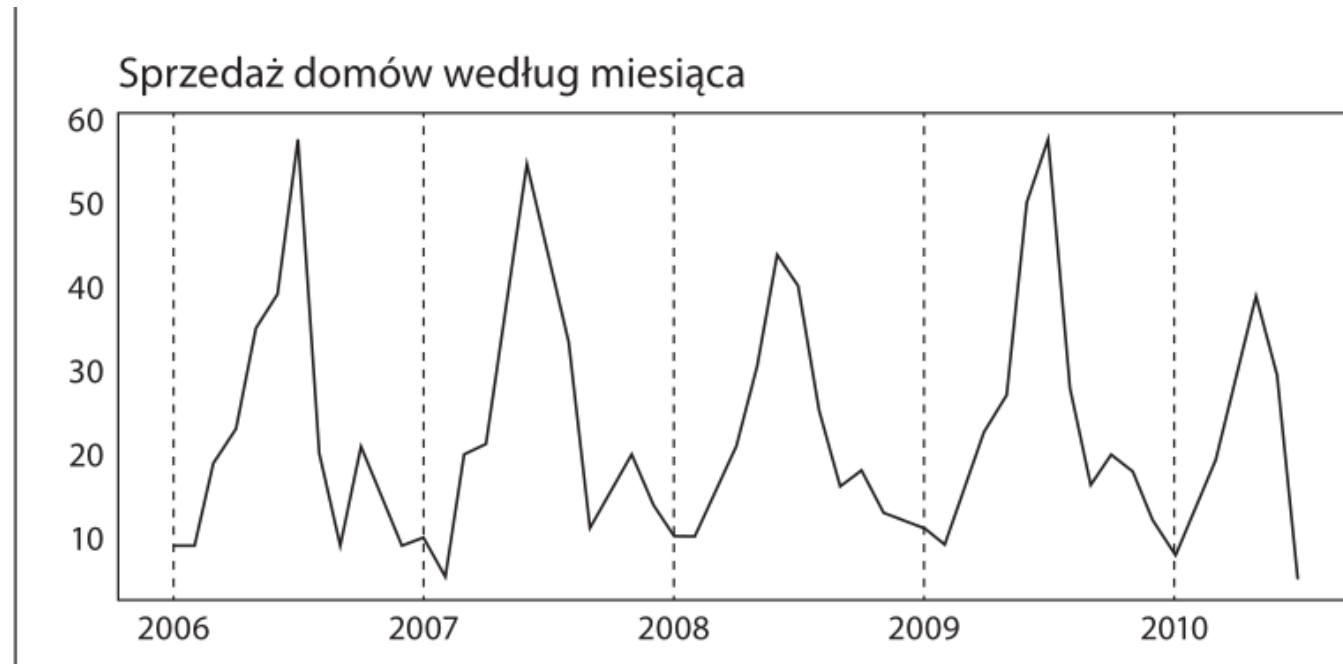
Wykres słupkowy

- Zlicza dane kategoryczne.
- Niemal wszystkie domy mają jednakową wartość tej cechy.
- Nie wpłynie ona na cenę sprzedaży.



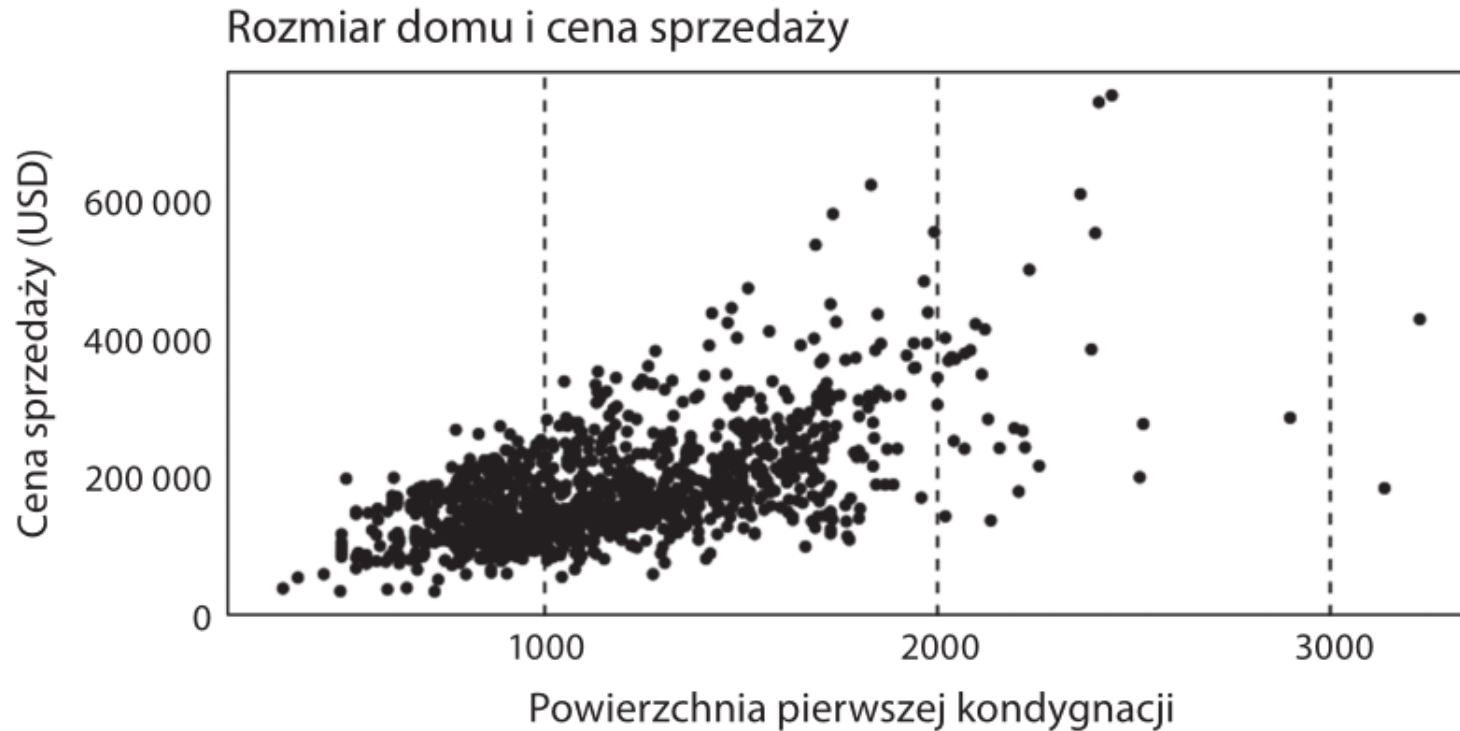
Wykres liniowy

- Liczba domów sprzedanych w poszczególnych miesiącach.
- Prezentuje on wzrosty sprzedaży latem i spadki zimą — przykład sezonowości.
- Wykresy liniowe pomagają dostrzegać takie trendy.



Wykres punktowy

- Domy według ich rozmiaru vs cena.
- Większe domy zwykle sprzedają się za większe kwoty

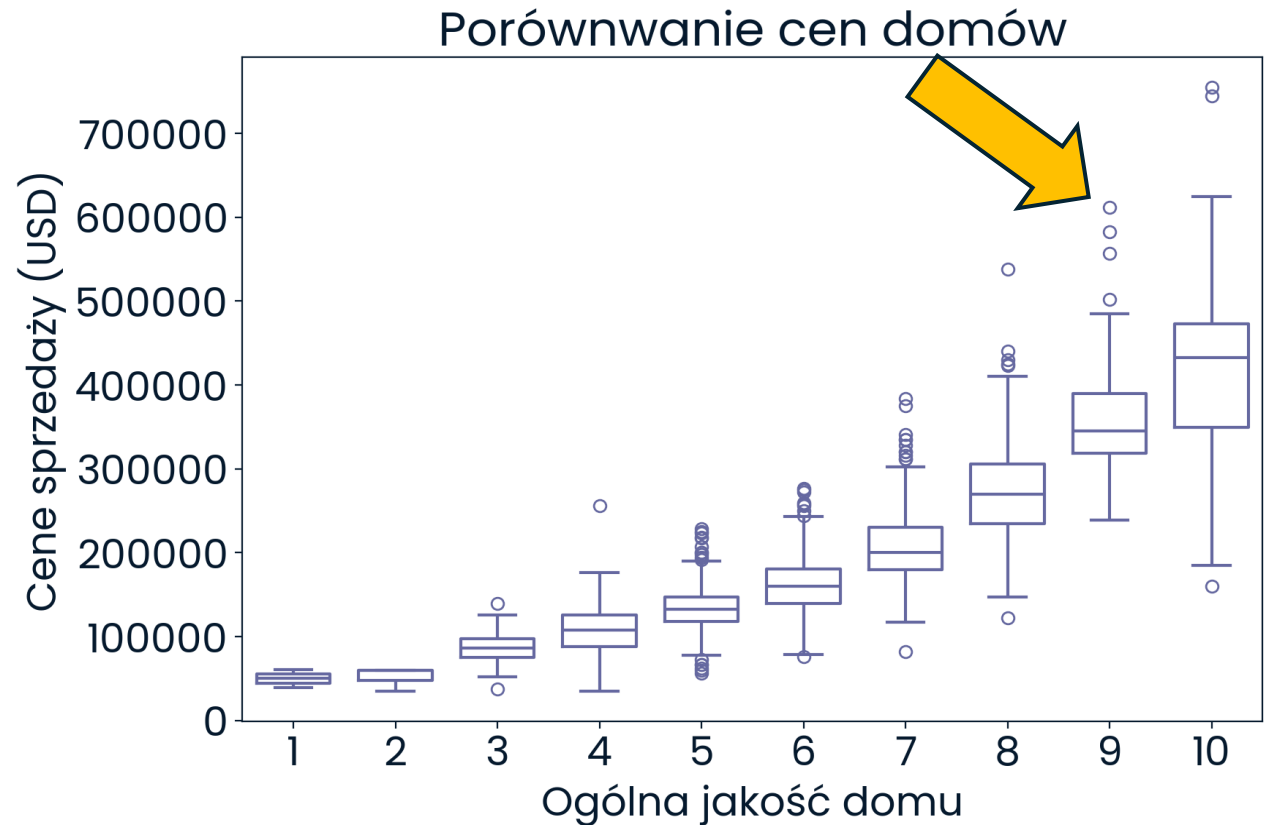


Wartości odstające i brakujące

- W każdym zbiorze danych znajdują się anomalie, wartości odstające i wartości brakujące.
- Sposób ich traktowania ma znaczenie.

Przykład

- Jednak samo to, że jakaś grafika klasyfikuje pewne punkty jako „wartości odstające”, nie oznacza jeszcze, że można automatycznie usunąć te punkty.



Zadanie EDA

- **Źródło:** <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>
- **Kontekst:** Zbiór zawiera dane o sprzedaży domów w Ames, Iowa (USA). Znajduje się w nim około 80 zmiennych opisujących różne aspekty nieruchomości - od fizycznych cech (powierzchnia, liczba pokoi) po lokalizację i jakość wykończenia.
- **Cel zadania:** Przeprowadzić eksploracyjną analizę danych (EDA), odkryć interesujące zależności i wizualnie przedstawić najważniejsze wnioski.

Pytania naprowadzające

- **Jaka jest zmienna docelowa** (target variable), którą chcemy przewidywać?
- **Jakie rodzaje zmiennych** masz w zbiorze? (numeryczne, kategoryczne, tekstowe?)

Do zrobienia:

- Wyświetl podstawowe statystyki opisowe
- Sprawdź typy danych wszystkich kolumn
- Zrobienie wykresów takich jak na poprzednich slajdach

Analiza zmiennej docelowej

- Jaki jest rozkład cen? (histogram)
- Czy są outliery (wartości odstające)?
- Jaka jest średnia/mediana ceny?
- Czy ceny są normalnie rozłożone? (wskazówka: czy histogram jest symetryczny?)

Stwórz **histogram** cen domów

- Poeksperymentuj z liczbą bins (10, 30, 50?)
- Co widzisz? Czy rozkład jest skośny?

Stwórz **box plot** dla cen

- Czy widzisz outliersy?
- Co mogą oznaczać te bardzo drogie domy?

Policz podstawowe statystyki:

min, max, średnia, mediana, Q1, Q3

Co wpływa na cenę domu?

Hipoteza do sprawdzenia: *"Większe domy są droższe"*

- Znajdź zmienne związane z powierzchnią (wskazówka: poszukaj słów "Area", "SF" w nazwach kolumn)
- Stwórz **scatter plot**: powierzchnia mieszkalna vs cena
 - Czy jest wyraźna zależność?
 - Czy jest liniowa?
 - Czy są jakieś dziwne obserwacje?

Dodatkowe pytania:

- Czy powierzchnia działki (LotArea) ma taki sam wpływ jak powierzchnia domu?
- Spróbuj stworzyć scatter plot dla 2-3 różnych typów powierzchni - który ma najsilniejszy związek z ceną?