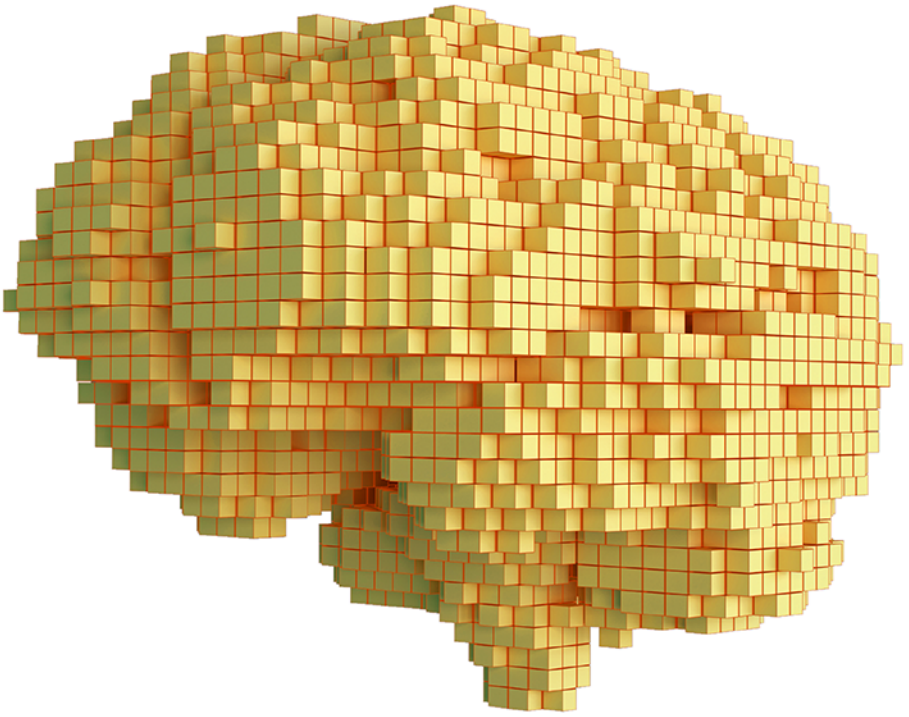
ALEX J. GUTMAN, JORDAN GOLDMEIER



ANALITYK DANYCH

PRZEWODNIK PO DATA SCIENCE, STATYSTYCE I UCZENIU MASZYNOWYM

Helion

Tytuł oryginału: Becoming a Data Head: How to Think, Speak and Understand Data Science, Statistics and Machine Learning

Tłumaczenie: Grzegorz Werner

ISBN: 978-83-289-0215-2

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/dascbi
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

ROZDZIAŁ 5

Eksploruj dane

„Jeśli powiesz specjaliście data science, żeby poszedł łowić ryby…
to zasługujesz na to, co otrzymasz, czyli kiepską analizę"[1].

— Thomas C. Redman, „Data Doc",
współpracownik „Harvard Business Review"

Projekty związane z danymi nigdy nie są takie proste, jakie wydają się podczas prezentacji w sali posiedzeń zarządu. Interesariusze zwykle widzą dopieszczoną prezentację PowerPointa, która przebiega według sztywno ustalonego skryptu, od pytania, poprzez dane, do odpowiedzi. Tym, czego nie widać w tej historii, są jednak wszystkie idee, które nie przeszły przez sito — ważne decyzje i założenia, które po drodze podjął zespół ds. danych, aby uzyskać odpowiedź. 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). Został sformułowany przez statystyka Johna Tukeya w latach 70. jako sposób na wstępne zrozumienie danych poprzez zbiorcze statystyki i wizualizacje przed zastosowaniem bardziej skomplikowanych metod[2]. Tukey postrzegał EDA jako pracę detektywistyczną. W danych ukryte są wskazówki, a właściwa eksploracja może zasugerować

[1] Cytat autorstwa Amy Gallo zaczerpnięty z *Understand Regression Analysis*, rozdział 10. przewodnika *HBR Guide to Data Analytics Basics for Managers*, HBR Guide Series.

[2] J.W. Tukey, *Exploratory data analysis*, 1977, vol. 2, s.131 – 160.

następne kroki. W rzeczywistości EDA jest kolejnym sposobem „polemizowania" z danymi. Jest to fundamentalna część pracy z danymi, która wyznacza i zmienia kierunek projektu w zależności od dokonywanych odkryć.

EKSPLORACYJNA ANALIZA DANYCH I TY

Eksploracyjna analiza danych 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. Po prostu 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.

Dlatego w tym rozdziale prezentujemy EDA jako ciągły proces, który należy do obowiązków każdego speca od danych, bez względu na to, czy bezpośrednio pracuje z danymi, czy jest członkiem zarządu firmy. Nauczysz się, jakie zadawać pytania i na co zwracać uwagę podczas eksplorowania danych.

Czy jesteś menedżerem, czy liderem?

Jeśli jesteś interesariuszem, menedżerem lub ekspertem merytorycznym, staraj się być dostępny dla swojego zespołu ds. danych. Prowadź otwarty dialog i oczekuj iteracji. Współpracuj z zespołem, aby ustalić poprawne założenia. Nie pozwól, aby zespół udał się na ryby bez odpowiedniego kontekstu biznesowego. W przeciwnym razie może podążyć ścieżkami, które mają większy sens statystyczny niż praktyczny. Jedno błędne założenie może zagrozić wszystkiemu, co nastąpi potem.

Rozumiemy, że menedżerowie nie mogą angażować się w szczegóły projektu w takim samym stopniu jak osoby bezpośrednio pracujące z danymi. Jest tu jednak miejsce na poprawę. Nie musisz stosować mikrozarządzania. Po prostu nie możesz ignorować[3].

PRZYJMIJ NASTAWIENIE EKSPLORACYJNE

Dziesiątki dostępnych narzędzi i języków programowania mogą pomóc zespołom ds. danych szybko i niedrogo eksplorować dane poprzez zbiorcze statystyki i wizualizacje. Nie należy jednak myśleć o EDA jak o przyborniku z narzędziami

[3] Żeby wszystko było jasne, interesariusze powinni powstrzymać się od mikrozarządzania. Konieczny jest pewien stopień zaufania między pracownikami biznesowymi a zespołami ds. danych.

albo liście kontrolnej. Jest to raczej nastawienie mentalne wplecione w każdą fazę pracy z danymi, które możesz przyjąć nawet bez zaplecza analitycznego.

Pytania naprowadzające

Aby pomóc Ci w przyjęciu nastawienia eksploracyjnego, przeprowadzimy Cię przez krótki scenariusz oparty na popularnym zbiorze danych utworzonym do celów edukacyjnych: *Ames Housing Data*[4]. Zapewni Ci to wgląd w proces EDA.

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?

Teraz nakreślimy ramy naszego scenariusza i omówimy każde z tych trzech pytań, aby wyjaśnić, dlaczego warto je zadać, i wskazać potencjalne problemy, które możesz napotkać.

Scenariusz

Pracujesz w start-upie działającym na rynku nieruchomości i musisz zwiększyć ruch w swojej witrynie. Trudno jednak odciągnąć klientów od technologicznych gigantów, takich jak amerykański serwis Zillow.com. Jego słynne narzędzie do wyceny nieruchomości, Zestimate®, przyciąga ludzi (i zyski) do marki Zillow[5]. Aby skutecznie konkurować z potentatami, Twoja firma potrzebuje własnego narzędzia predykcyjnego. Otrzymałeś(-łaś) więc zadanie zbudowania *modelu*, który przyjmuje informacje o domu jako *dane wejściowe* i generuje szacowaną cenę sprzedaży jako *dane wyjściowe*.

Na dobry początek szef wysyła Ci zbiór danych. Ma on 80 kolumn opisujących różne aspekty setek domów mieszkalnych sprzedanych w Ames w stanie Iowa w latach 2006 – 2011.

Tak duża ilość danych może wydawać się przytłaczająca. Jednak wymienione poprzednio pytania pomogą Ci przygotować wstępny plan pracy z danymi.

Omówmy je po kolei.

---

[4] D. De Cock, *Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project*, „Journal of Statistics Education" 2013, vol. 19(3). Dane możesz pobrać pod adresem *www.kaggle.com/c/house-prices-advanced-regression-techniques*.

[5] Firma Zillow traktuje swoje narzędzie Zestimate® bardzo poważnie. W 2019 roku przyznała milion dolarów nagrody zespołowi specjalistów data science za zwiększenie dokładności szacunków Zestimate®: *venturebeat.com/2019/01/30/zillow-awards-1--million-to-team-that-reduced-home-valuation- algorithm-error-to-below-4*.

CZY DANE MOGĄ ODPOWIEDZIEĆ NA PYTANIE?

Choć możesz odczuwać pokusę, aby natychmiast przepuścić dane przez obecnie modny algorytm (na przykład uczenie głębokie opisane w rozdziale 12.), najpierw musisz zapytać: „Czy dane mogą odpowiedzieć na pytanie?". A żeby znaleźć odpowiedź, często wystarczy po prostu przyjrzeć się danym.

Określ oczekiwania i użyj zdrowego rozsądku

Powinieneś (powinnaś) dość dobrze wiedzieć, jakie informacje są potrzebne do oszacowania ceny sprzedaży domu: rozmiar, liczba sypialni, liczba łazienek, rok budowy itd. Są to popularne cechy, które potencjalni kupujący wyszukują w Twojej witrynie. Przewidywanie ceny sprzedaży bez tych informacji nie wydawałoby się rozsądne.

Nazwy kolumn i typy danych są widoczne od razu po otwarciu pliku. Obecne są tam zdroworozsądkowe informacje, których oczekujesz, a także pomocne dane porządkowe (ogólna jakość domu w skali 1 – 10, przy czym 10 to jakość „doskonała"), dane nominalne (dzielnica) i wiele innych cech. Dane przeszły zatem wstępny test.

Następnie powinieneś (powinnaś) zbadać wartości zmiennych. Czy obejmują one scenariusze, które chcesz analizować? Jeśli na przykład odkryjesz, że zmienna „Typ nieruchomości" uwzględnia tylko domy jednorodzinne, ale nie apartamenty, bliźniaki albo mieszkania, Twój model będzie miał ograniczony zakres w porównaniu z tym, którego używa Zillow. Zestimate® może przewidzieć cenę sprzedaży mieszkania, ale jeśli nie masz historycznych danych dotyczących mieszkań, Twoja firma nie będzie mogła precyzyjnie szacować ich cen.

Lekcja: unikaj wypraw na ryby, przed którymi ostrzegał Cię cytat na początku tego rozdziału. Upewnij się, że dane mają sens z perspektywy tego, jak będą wykorzystywane.

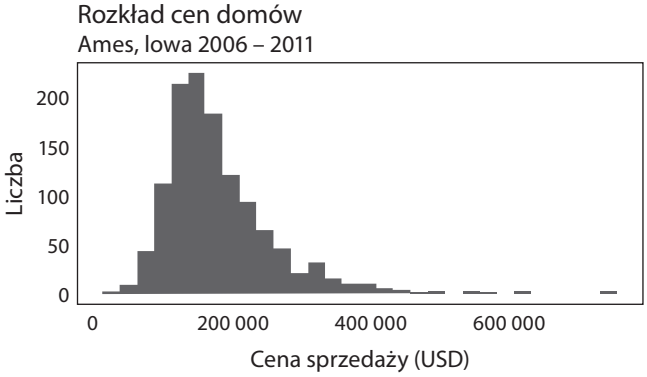
Czy wartości mają intuicyjny sens?

Możesz użyć oprogramowania, aby wygenerować mnóstwo zbiorczych statystyk. Twoim zadaniem jest umieścić dane w kontekście. Sprawdź, czy zbiorcze statystyki odpowiadają Twojemu intuicyjnemu rozumieniu problemu. Równie ważnym elementem EDA są wizualizacje — wykorzystaj je do odkrywania anomalii i innych dziwactw w danych.

Powtórka z wizualizacji danych

Omówmy kilka przykładów EDA z histogramami oraz wykresami skrzynkowymi, słupkowymi i punktowymi. Możesz pominąć tę ramkę, jeśli rozumiesz te wykresy i wiesz, co mogą Ci podpowiedzieć.

Możesz poznać kształt albo rozkład ciągłych danych liczbowych poprzez przyjrzenie się histogramowi. Rozważ histogram cen sprzedaży pokazany na rysunku 5.1. Widać na nim przykład około 125 domów w przedziale 200 000 dol. i długi „ogon" po prawej stronie, który reprezentuje najdroższe domy. Ogon ten przeciąga średnią cenę sprzedaży (181 000 dol.) poza medianę (163 000 dol.). Kilka drogich domów sprawia, że średnia jest większa niż mediana.

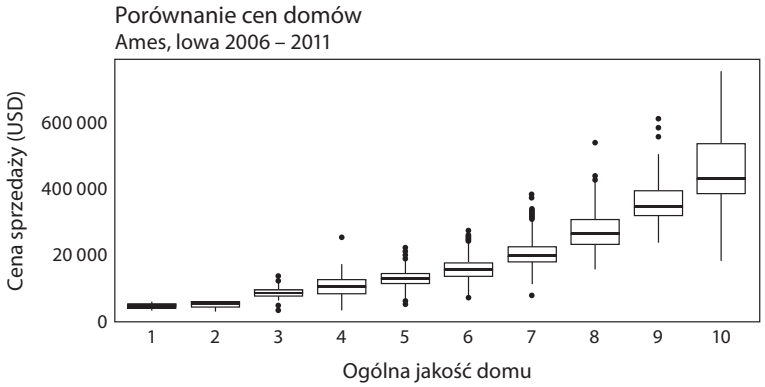


RYSUNEK 5.1. Histogram pokazujący kształt cen sprzedaży

Histogramy pomagają dostrzegać anomalie. Gdybyś zobaczył(a) wartości ujemne (ktoś otrzymuje zapłatę za kupienie domu?) albo przedziały z bardzo dużymi liczbami na skrajach rysunku 5.1, co zdarza się często, kiedy dane są „obcięte" (na przykład każda wartość powyżej 500 000 dol. jest wprowadzana jako 500 000 dol.), powinieneś (powinnaś) zadać kilka pytań.

Wykresy skrzynkowe[6] można wykorzystać do porównywania danych w kilku grupach. Na rysunku 5.2 pokazano wykres skrzynkowy dla ocen jakości domu, gdzie 1 to jakość bardzo niska, a 10 to doskonała.

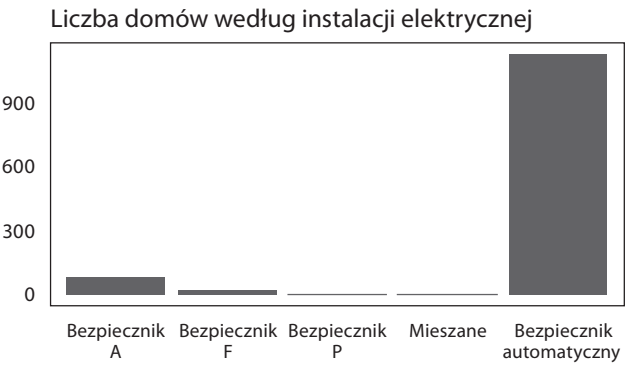
[6] Wykresy skrzynkowe nazywa się również wykresami typu „skrzynka i wąsy". „Skrzynka" zawiera środkową połowę danych (wartości między 25. a 75. centylem), linia w środku to mediana, a „wąsy" pokazują zakres pozostałych punktów danych. Kropki pod wąsami to potencjalne wartości odstające.



RYSUNEK 5.2. Używanie wykresów skrzynkowych do porównywania cen przy różnych ocenach jakości

Związek między ogólną jakością a cenami domów wydaje się zgodny z intuicją. Domy wyższej jakości zwykle sprzedają się za wyższe kwoty. Dostrzegamy dom za 200 000 dol. z jakością ocenioną na 10 (dolny koniec linii), ale rozsądne wydaje się założenie, że sprzedano go za mniej niż inne idealne „dziesiątki" ze względu na inne czynniki. Jest to dokładnie ten rodzaj informacji, które powinien sprawdzać każdy, kto pracuje z danymi.

Wykresy słupkowe, takie jak pokazany na rysunku 5.3, zliczają dane kategoryczne.



RYSUNEK 5.3. Wykres słupkowy pokazujący liczby domów według instalacji elektrycznej

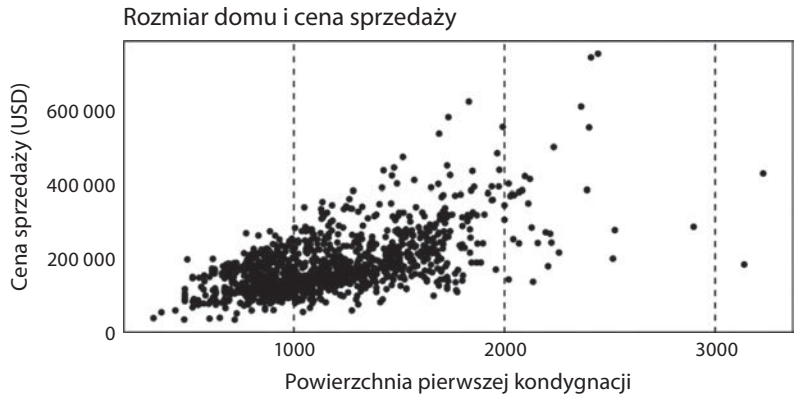
Nie wszystkie obrazy są interesujące na pierwszy rzut oka. Warto jednak obejrzeć różne wizualizacje choćby po to, aby potwierdzić (a może zakwestionować) odpowiedź na poprzednie pytanie — czy dane mają intuicyjny sens? Na przykład rysunek 5.3 pokazuje, że niemal wszystkie domy mają

jednakową wartość tej cechy. Dla Twoich celów informacja ta jest jednak przydatna. Ponieważ większość domów ma jednakową wartość tej zmiennej, prawdopodobnie nie wpłynie ona znacząco na ceny sprzedaży domów.

Na rysunku 5.4 pokazano wykres liniowy z liczbą domów sprzedanych w poszczególnych miesiącach. Prezentuje on wzrosty sprzedaży latem i spadki zimą — przykład *sezonowości*. Wykresy liniowe pomagają dostrzegać takie trendy.

RYSUNEK 5.4. Wykres liniowy pokazujący liczbę domów sprzedanych w poszczególnych miesiącach

Następnie zbadajmy wykres punktowy pokazujący domy według ich rozmiaru (powierzchni pierwszej kondygnacji) i ceny sprzedaży (rysunek 5.5).



RYSUNEK 5.5. Wykres punktowy pokazujący powierzchnię i cenę sprzedaży

Z rysunku 5.5 wyłania się interesujący wzorzec. Większe domy zwykle sprzedają się za większe kwoty. Oczywiście, ta reguła nie zawsze się sprawdza. Czasem mniejsze domy kosztują więcej niż większe. Zawsze jest jakaś zmienność, ale ogólny trend jest oczywisty. Ponieważ zaś próbujemy przewidywać ceny sprzedaży, powierzchnia domu wydaje się bardzo cenną informacją.

Był to tylko przedsmak informacji i spostrzeżeń, które możesz szybko poczynić poprzez naniesienie danych na wykres. Gdybyś chciał(a) dowiedzieć się więcej o wizualizacji w eksploracji danych, polecamy następujące tytuły:

- Stephen Few, *Now You See it: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press, 2009.
- Edward Tufte, *The Visual Display of Quantitative Information*, Edward Tufte Graphics Press, 2011.

Uważaj! Wartości odstające i brakujące

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

Na przykład wykres skrzynkowy z rysunku 5.2 wykorzystuje praktyczne reguły, aby oznaczyć kilka punktów danych jako potencjalne wartości odstające. Jednak samo to, że jakaś grafika klasyfikuje pewne punkty jako „wartości odstające", nie oznacza jeszcze, że możesz wyłączyć krytyczne myślenie i automatycznie usunąć te punkty, zakładając, że nie będą użyteczne. Zillow na pewno nie usuwa przydatnych informacji ze swoich zbiorów danych tylko dlatego, że wizualizacja opisuje je jako odstające. Zwróć uwagę na kontekst danych — domy, które kosztują znacznie więcej niż większość innych, to znana i częsta cecha danych dotyczących nieruchomości. Przypomnij sobie lekcje z poprzedniego rozdziału. Powinieneś (powinnaś) mieć dobre uzasadnienie biznesowe, aby usunąć wartości odstające. Czy masz je w tym przypadku?

A co z brakującymi wartościami? Czy brakująca wartość zmiennej „Rozmiar piwnicy" oznacza, że dom ma piwnicę, ale jej powierzchnia jest nieznana? Czy może nie ma piwnicy i wartość powinna wynosić 0?

Jeśli wydaje się, że mnożymy problemy, to właśnie taki był nasz zamiar. Osoby pracujące z danymi podejmują setki takich drobnych decyzji podczas każdego projektu. Ich łączny efekt może być znaczący. Pozostawieni samym sobie — i pozbawieni dostępu do wiedzy dziedzinowej — pracownicy zajmujący się danymi mogą po trochu je oskubywać, usuwając trudne i zniuansowane przypadki, aż dane staną się zbyt oddalone od rzeczywistości, żeby były użyteczne. Właśnie dlatego wszyscy, łącznie z menedżerami, powinni naprawdę rozumieć, co robią zespoły ds. danych.

CZY ODKRYLIŚCIE JAKIEŚ ZWIĄZKI?

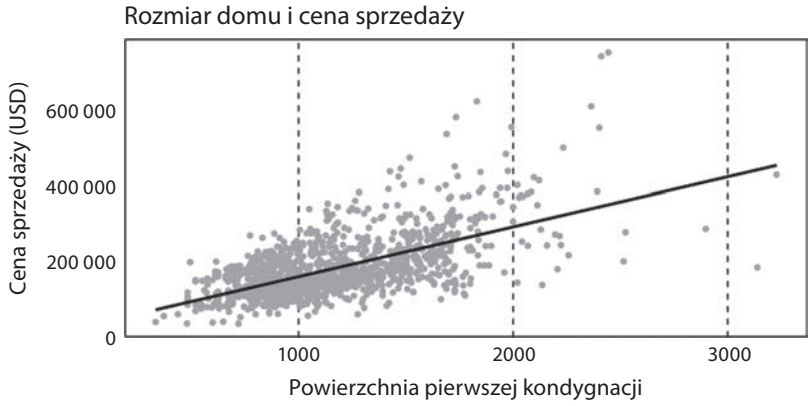
Wstępna analiza danych mieszkaniowych z wykorzystaniem zbiorczych statystyk i wizualizacji wydaje się zachęcająca i uważasz, że danych rzeczywiście będzie można użyć do zbudowania modelu przewidującego ceny sprzedaży, więc przechodzisz do następnego pytania: „Czy odkryliście jakieś związki?".

Wizualizacja danych dała Ci pierwsze wskazówki: wyższa ogólna jakość i większy metraż są, jak można było się spodziewać, związane z wyższymi cenami sprzedaży. Właśnie takich informacji zwrotnych szukasz w danych. Związki wydają się sensowne, a zmienne, które naniosłeś(-łaś) na wykres, pomogą Ci zbudować model do przewidywania ceny sprzedaży. Jakie inne zmienne mają związek z ceną sprzedaży?

W tym momencie na interesujące wzorce i relacje w danych mogą nakierować Cię statystyki zbiorcze, ponieważ generowanie każdego możliwego wykresu punktowego może nie być praktyczne. Relację znalezioną na wykresach punktowych można sprowadzić do statystyki zbiorczej zwanej **korelacją**, która wskazuje na związek między dwiema zmiennymi liczbowymi (choć go nie dowodzi).

Korelacja

Korelacja jest miarą tego, jak powiązane są dwie zmienne. Najczęstszym typem korelacji używanym w biznesie jest **współczynnik korelacji Pearsona**, statystyka o wartości od –1 od 1, która mierzy liniowy związek (wyrażony zwykłymi liniami prostymi) między parami liczb na wykresie punktowym. Korelacja może być dodatnia, co oznacza, że zwiększenie jednej zmiennej powoduje wzrost drugiej: większe domy sprzedają się za wyższe kwoty. Może też być ujemna: cięższe samochody przejeżdżają mniej kilometrów na tej samej ilości benzyny. Korelacja między rozmiarem domu a ceną sprzedaży, pokazana na rysunku 5.6, wynosi 0,62. Im „ciaśniej" skupione są punkty wokół linii trendu, tym wyższa korelacja[7].



RYSUNEK 5.6. Metraż i cena sprzedaży mają korelację równą 0,62, która mierzy, jak ciasno skupiają się punkty wokół linii trendu

[7] Korelacja nie oznacza „stromizny". Wykresy dwóch ściśle skorelowanych zmiennych mogą wydawać się niemal płaskie (choć niezupełnie poziome).

Korelacja może Ci pomóc na dwa sposoby. Po pierwsze, znalezienie zmiennych skorelowanych z ceną sprzedaży pomogłoby ją przewidywać. Po drugie, korelacja może ograniczyć redundancję w danych, ponieważ dwie ściśle skorelowane zmienne zawierają mniej więcej te same informacje. Wyobraź sobie, że masz w danych dwie kolumny: powierzchnię domu w stopach kwadratowych i w metrach kwadratowych. Te dwie zmienne są idealnie skorelowane; do analizy potrzebna jest tylko jedna.

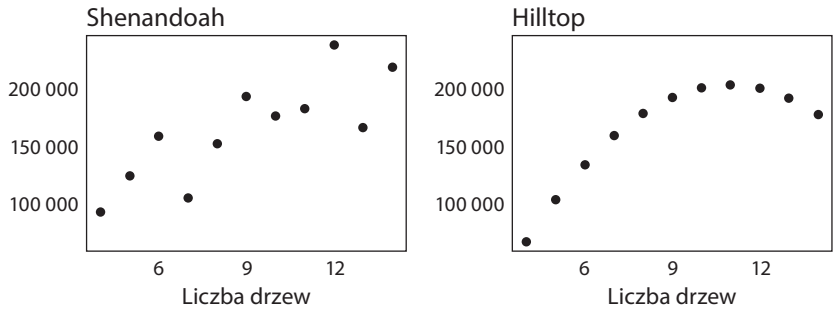
Choć większość z nas wie co nieco o korelacji i często posługuje się tą miarą, bywa ona zwodnicza. Sprawdźmy dlaczego.

Uważaj! Błędne interpretowanie korelacji

Często zapomina się, że korelacja jest miarą trendu liniowego, a nie wszystkie trendy są liniowe.

Przypuśćmy na przykład, że analizujesz dwie dzielnice w zbiorze danych mieszkaniowych, każdą z 11 domami. Szybkie obliczenia ujawniają, że liczba drzew na działce jest ściśle skorelowana z ceną sprzedaży w tych dzielnicach. Korelacja wynosi 0,8: działki z większą liczbą drzew sprzedają się drożej.

Jednak wizualna kontrola danych ujawnia coś nieoczekiwanego. Na rysunku 5.7 dane po lewej stronie ukazują obraz, którego spodziewamy się po wysokiej korelacji: liniowy trend z nieco rozrzuconymi punktami danych. Ale wykres po prawej stronie pokazuje, że liczba drzew jest związana ze wzrostem ceny *tylko do pewnego punktu* (11 drzew). Następnie linia opada. Może w dzielnicy Hilltop na niektórych działkach rośnie tyle drzew, że przeszkadzają one w pielęgnowaniu trawnika.



RYSUNEK 5.7. Dwa zbiory danych z korelacją równą 0,8

Pora uczciwie przyznać: dane pokazane na rysunku 5.7 nie pochodzą z badanego przez nas zbioru danych Ames, ale z popularnego zestawu Anscombe's Quartet[8] składającego się z czterech zbiorów danych z iden-

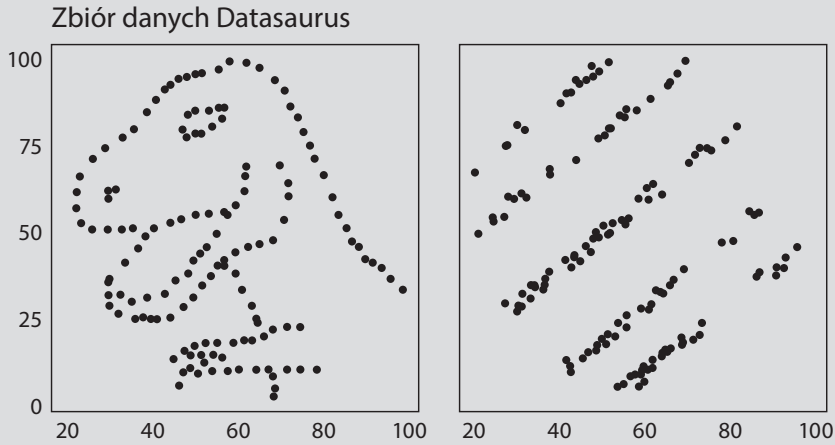
[8] F.J. Anscombe, *Graphs in statistical analysis*, „The American Statistician" 1973, vol. 27(1), s. 17 – 21. Pomnożyliśmy zmienną zależną przez 22 000, aby utworzyć realistyczne przykłady cen sprzedaży.

tycznymi statystykami zbiorczymi, ale wyraźnie różnymi wizualizacjami. (Pokazaliśmy tylko dwie i dostosowaliśmy dane, aby pozostać przy temacie sprzedaży nieruchomości).

Lekcja: używaj wizualizacji, aby zweryfikować godne uwagi korelacje w danych, ponieważ liniowy trend zidentyfikowany przez korelację może nie opowiadać pełnej historii.

Nieskorelowane, ale wciąż interesujące

Na rysunku 5.8 pokazano dwa wykresy z niemal identycznymi, bliskimi zera współczynnikami korelacji. Nie oznacza to jeszcze, że nie ma w nich nic interesującego. I choć nie napotkasz wielu „danozaurów", jak ten pokazany na lewym wykresie, możesz trafić na scenariusz z prawego wykresu: pięć grup skorelowanych liniowo danych, które jako całość w ogóle nie wykazują korelacji. Jest to tzw. paradoks Simpsona, o którym dowiesz się więcej w rozdziale 13.



RYSUNEK 5.8. Danozaur: możesz sam pobrać i zbadać dane[9]. Podobnie jak w przypadku Anscombe's Quartet, oba pokazane tu zbiory danych mają identyczne statystyki zbiorcze

[9] Danozaur został stworzony przez Alberto Caira, a dane są dostępne na GitHubie: *github.com/lockedata/datasauRus*.

Uważaj! Korelacja nie implikuje przyczynowości

Prawdopodobnie słyszałeś(-łaś) już wcześniej powiedzenie: „Korelacja nie implikuje przyczynowości”[10]. Warto je jednak tu powtórzyć, ponieważ jest często ignorowane, a nawet źle rozumiane.

Kiedy dwie zmienne są skorelowane, nawet bardzo ściśle, nie oznacza to, że jedna powoduje drugą. Ludzie jednak często wpadają w tę pułapkę, próbując zbudować jakąś narrację za każdym razem, kiedy dwie zmienne poruszają się razem. Istnieją typowe niemądre przykłady, których używają statystycy, aby pokazać, że korelacja nie oznacza przyczynowości. Sprzedaż lodów jest skorelowana z atakami rekinów (jedno i drugie rośnie w miesiącach letnich). Rozmiar buta jest skorelowany z umiejętnością czytania (jedno i drugie rośnie z biegiem czasu). Jednak sugerowanie, że zmniejszenie sprzedaży lodów ograniczyłoby liczbę ataków rekinów albo że kupno większych butów pomogłoby w czytaniu to oczywiście żart. W grę wchodzą inne czynniki — temperatura zewnętrzna w przykładzie z lodami, wiek w przykładzie z rozmiarem buta — które odgrywają oczywistą rolę w złudnym związku.

Kiedy jednak korelacje nie pojawiają się w kontekście żartów, a prawdziwe przyczyny nie są znane, mantra „korelacja nie implikuje przyczynowości” często idzie w zapomnienie.

Na przykład w danych dotyczących nieruchomości widać, że miary edukacyjnych wyników szkół są skorelowane z wartością domów. Czy oznacza to, że lepsze szkoły powodują wzrost wartości domu? Można domniemywać, że dobre szkoły podnoszą atrakcyjność dzielnicy. A może przyczynowość działa w odwrotnym kierunku: wyższe ceny domów powodują poprawę edukacyjnych wyników szkół? Może większe przychody z podatków zapewniają szkołom więcej zasobów? A może przyczynowość działa w obu kierunkach, tworząc sprzężenie zwrotne? W większości przypadków po prostu tego nie wiemy. W grę wchodzi mnóstwo innych czynników i rzadko zdarza się, żebyś miał(a) wszystkie odpowiedzi w swoim zbiorze danych.

Bezpieczniej jest założyć, że „nie ma przyczynowości” między dwiema skorelowanymi zmiennymi, dopóki ktoś nie przeprowadzi eksperymentu, który jej dowiedzie. Nie należy jednak popadać w skrajność. Obaj autorzy widzieli, jak w środowiskach biznesowych, uniwersyteckich i medialnych zakładano przyczynowość tam, gdzie było to nieuzasadnione. Bywają też jednak przypadki, w których ważny związek jest natychmiast odrzucany jako rzekomo błędne założenie przyczynowości. (Przykład odrzucenia przyczynowości w sytuacji, w której nie należało tego robić, podajemy w ramce).

[10] Autorzy zastanawiali się nawet, czy można napisać książkę o danych i nie wspomnieć o tym, że „korelacja nie implikuje przyczynowości”. Jaki był wynik tej dyskusji, widzisz sam.

Palenie a rak płuc

Ronald A. Fisher, jeden z czołowych statystyków XX wieku, który zresztą był współtwórcą wielu technik opisywanych w tej książce, podchodził dość sceptycznie do badań wiążących palenie tytoniu z rakiem.

Fisher obawiał się o wpływ zmiennych zakłócających. Co by było, gdyby na przykład niektórzy ludzie byli genetycznie predysponowani do zachorowania na raka płuc i palili, aby złagodzić dolegliwości? Według Fishera we wczesnych badaniach nad zagrożeniami związanymi z używaniem tytoniu popełniono „błąd [...] starego rodzaju, przechodząc od korelacji do związku przyczynowego”[11].

A jednak obecnie wiemy, że ten związek jest niepodważalny. Choć powinniśmy wystrzegać się dostrzegania przyczynowości tam, gdzie jej nie ma, musimy również uważać, aby nie odrzucać jeszcze niedowiedzionych związków przyczynowych.

CZY ZNALEŹLIŚCIE W DANYCH NOWE MOŻLIWOŚCI?

EDA nie tylko pozwala lepiej zrozumieć dane i wyznaczyć drogę do rozwiązania problemu. Daje też szansę znalezienia w danych dodatkowych możliwości; nowych problemów, które mogą być cenne dla Twojej organizacji. Specjalista data science może dostrzec w zbiorze danych coś interesującego lub nietypowego, a następnie sformułować problem.

Nie będziesz jednak wiedział(a), czy ktokolwiek potrzebuje rozwiązania, które znalazłeś(-łaś), dopóki nie wykonasz czynności opisanych w rozdziale 1. „Na czym polega problem?”.

PODSUMOWANIE ROZDZIAŁU

Aby zostać specem od danych, musisz polubić proces eksploracyjnej analizy danych. Pozwoli Ci to:

- znaleźć klarowniejszą ścieżkę do rozwiązania problemu;
- doprecyzować pierwotny problem biznesowy zgodnie z ograniczeniami odnalezionymi w danych;
- zidentyfikować nowe problemy, które można rozwiązać z wykorzystaniem dostępnych danych;
- anulować projekt. Kiedy wyniki są niezadowalające, analiza EDA zapobiega traceniu czasu i pieniędzy na projekty, które zabrnęły w ślepą uliczkę.

[11] R.A. Fisher, *Cancer and smoking*, „Nature” 1958, vol. 182 (4635), s. 596.

Zademonstrowaliśmy proces EDA na przykładzie zbioru danych dotyczących nieruchomości (do którego wrócimy w rozdziale 9., aby wreszcie zbudować ten model, o którym mówiliśmy) i opisaliśmy przeszkody, które możesz napotkać po drodze.

W rozdziale tym założyliśmy, że możesz być częścią procesu EDA od początku do końca. Czasem jest to niemożliwe, zwłaszcza w przypadku dyrektorów wyższego szczebla, którzy nadzorują wiele projektów. Ale nieobecność we wczesnych fazach projektu nie zwalnia speców od danych z obowiązku przyjęcia eksploracyjnego nastawienia. Jeśli dołączasz do projektu pod koniec, zapytaj, dlaczego zespół ds. danych wybrał konkretny sposób analizy i jakie napotkał trudności. Możesz odkryć założenia, których nie poczynił(a)byś sam(a).

Data science? Odsiejesz piasek od złota!

Musisz spojrzeć prawdzie w oczy: epoka danych to nie tylko imponujące możliwości, ale również obietnice bez pokrycia. Firmy wdrażają rozwiązania, które mają je wyręczać w podejmowaniu decyzji. Menedżerowie zatrudniają analityków, którzy nimi nie są. Specjaliści w dziedzinie data science są zatrudniani w organizacjach, które nie są na nich gotowe. Dyrektorzy wysłuchują technicznego żargonu i udają, że go rozumieją. Efekt? Pieniądze idą w błoto.

Oto praktyczny przewodnik po nauce o danych w miejscu pracy. Dowiesz się stąd wszystkiego, co ważne na początku Twojej drogi jako danologa: od osobowości, z którymi przyjdzie Ci pracować, przez detale analizy danych, po matematykę stojącą za algorytmami i uczeniem maszynowym. Nauczysz się myśleć krytycznie o danych i otrzymanych wynikach, będziesz też inteligentnie o tym mówić. Jednym zdaniem: zrozumiesz dane i związane z nimi wyzwania na głębszym, profesjonalnym poziomie.

> ***To książka dla każdego, kto chce przestawić firmę na tory data science.***
>
> **Eric Weber**
> *kierownik ds. eksperymentów i badań metrycznych, Yelp*

Naucz się:

- myśleć statystycznie i rozumieć rolę zmienności w podejmowaniu decyzji
- zadawać właściwe pytania na temat statystyk i wyników analiz
- sensownie korzystać z rozwiązań uczenia maszynowego i sztucznej inteligencji
- unikać typowych błędów podczas pracy z danymi i ich interpretowania

Dr Alex J. Gutman jest adiunktem w Instytucie Technicznym Wojsk Lotniczych, specjalistą z zakresu data science i instruktorem biznesowym.

Jordan Goldmeier jest światowej klasy ekspertem w dziedzinie analityki i wizualizacji danych, a także autorem książek. Od sześciu lat otrzymuje nagrodę Excel MVP. Jest również wolontariuszem ratownictwa medycznego.

Helion
helion.pl
HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-289-0215-2

Cena: 69,00 zł

WILEY