

DAVID
ASBOTH

Jak ogarnąć
**TRUDNE
DANE?**

**Praktyczne podejście
profesjonalnego analityka**

Helion 

Tytuł oryginału: The Well-Grounded Data Analyst: Solve messy data problems like a pro

Tłumaczenie: Anna Mizerska

Projekt okładki: Studio Gravite / Olsztyn

Obarek, Pokoński, Pazdrijowski, Zaprucki

Materiały graficzne na okładce zostały wykorzystane za zgodą Adobe Stock.

ISBN: 978-83-289-3308-8

© Helion S.A. 2026.

Authorized translation of the English edition © 2025 Manning Publications.
This translation is published and sold by permission of Manning Publications,
the owner of all rights to publish and sell the same.

All rights reserved. No part of this book may be reproduced or transmitted
in any form or by any means, electronic or mechanical, including photocopying,
recording or by any information storage retrieval system, without permission
from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości
lub fragmentu niniejszej publikacji w jakiegokolwiek postaci jest zabronione.
Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie
książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie
praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi
bądź towarowymi ich właścicieli.

Autor oraz wydawca dołożyli wszelkich starań, by zawarte w tej książce informacje
były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich
wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych
lub autorskich. Autor oraz wydawca nie ponoszą również żadnej odpowiedzialności
za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Drogi Czytelniku!

Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres

helion.pl/user/opinie/jakotd

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: helion.pl (księgarnia internetowa, katalog książek)

Printed in Poland.

- [Kup książkę](#)
- [Poleć książkę](#)
- [Oceń książkę](#)

- [Księgarnia internetowa](#)
- [Lubię to! » Nasza społeczność](#)

Spis treści

<i>Przedmowa</i>	13
<i>Wstęp</i>	15
<i>Podziękowania</i>	17
<i>O książce</i>	19
<i>O autorze</i>	23
1. Łączenie teorii z praktyką w nauce o danych	25
1.1. Zestaw narzędzi analityka danych	26
1.2. Podejście zorientowane na wyniki	27
1.2.1. Zrozum problem	29
1.2.2. Zaczynij od końca	30
1.2.3. Poszukaj dodatkowych źródeł	31
1.2.4. Pozyskaj dane	31
1.2.5. Wykonaj pracę	31
1.2.6. Przedstaw minimalną satysfakcjonującą odpowiedź	32
1.2.7. W razie potrzeby iteruj	32
1.3. Struktura projektu	33
Podsumowanie	35
2. Praca z danymi geograficznymi	36
2.1. Projekt 1. Określenie lokalizacji klientów	37
2.1.1. Słownik danych	38
2.2. Przykładowe rozwiązanie: odnajdywanie Londynu	39
2.2.1. Przygotowanie do sukcesu	39
2.2.2. Tworzenie pierwszej wersji rozwiązania	40
2.2.3. Podsumowanie i kolejne kroki	54
2.3. Jak czytać pozostałą część książki?	56
Podsumowanie	57

3.	<i>Modelowanie danych</i>	58
3.1.	Znaczenie modelowania danych	59
3.1.1.	<i>Typowe zadania w modelowaniu danych</i>	60
3.2.	Projekt 2. Kim są Twoi klienci?	63
3.2.1.	<i>Założenia projektu</i>	63
3.2.2.	<i>Słownik danych</i>	64
3.2.3.	<i>Oczekiwane rezultaty</i>	66
3.2.4.	<i>Niezbędne narzędzia</i>	66
3.3.	Planowanie podejścia do modelowania danych klientów	67
3.3.1.	<i>Zastosowanie procesu zorientowanego na wyniki do modelowania danych</i>	67
3.3.2.	<i>Pytania wartę uwagi</i>	70
3.4.	Przykładowe rozwiązanie: identyfikacja klientów na podstawie danych transakcyjnych	71
3.4.1.	<i>Opracowanie planu działania</i>	71
3.4.2.	<i>Odkrywanie, pozyskiwanie i łączenie danych z wielu źródeł</i>	71
3.4.3.	<i>Rozwiązywanie problemu encji w celu usunięcia duplikatów</i>	91
3.4.4.	<i>Wnioski i zalecenia</i>	102
3.5.	Podsumowanie modelowania danych	104
3.5.1.	<i>Umiejętności modelowania danych przydatne w każdym projekcie</i>	105
	Podsumowanie	105
4.	<i>Wskaźniki</i>	106
4.1.	Znaczenie dobrze zdefiniowanych wskaźników	107
4.2.	Projekt 3. Definiowanie precyzyjnych wskaźników po to, by podejmować lepsze decyzje	108
4.2.1.	<i>Opis problemu</i>	108
4.2.2.	<i>Słownik danych</i>	109
4.2.3.	<i>Oczekiwane rezultaty</i>	110
4.2.4.	<i>Wymagane narzędzia</i>	110
4.3.	Zastosowanie metody zorientowanej na wyniki do różnych definicji wskaźników	111
4.3.1.	<i>Pytania wartę uwagi</i>	114
4.4.	Przykładowe rozwiązanie: identyfikacja najlepiej sprzedających się produktów	114
4.4.1.	<i>Łączenie i analiza danych o produktach</i>	114
4.4.2.	<i>Obliczanie wskaźników na poziomie produktu</i>	128
4.4.3.	<i>Znajdowanie najlepszych produktów z użyciem zdefiniowanych wskaźników</i>	134

4.5. Podsumowanie rozdziału o wskaźnikach	147
4.5.1. Umiejętności definiowania lepszych wskaźników dla dowolnego projektu	148
Podsumowanie	149
5. Nietypowe źródła danych	150
5.1. Szukanie nowych źródeł danych	151
5.1.1. Kwestie do rozważenia przy korzystaniu z nowych zbiorów danych	151
5.2. Projekt 4. Analiza trendów w branży filmowej z użyciem danych z plików PDF	153
5.2.1. Opis problemu	153
5.2.2. Słownik danych	154
5.2.3. Oczekiwane rezultaty	155
5.2.4. Wymagane narzędzia	156
5.3. Zastosowanie metody zorientowanej na wyniki do ekstrakcji danych z plików PDF	156
5.4. Przykładowe rozwiązanie: wpływ ograniczeń związanych z COVID-19 na przemysł filmowy	158
5.4.1. Analiza dostępnych danych	159
5.4.2. Wyodrębnianie danych z plików PDF	164
5.4.3. Analiza danych wyodrębnionych z plików PDF	173
5.4.4. Wnioski i zalecenia dotyczące projektu	186
5.5. Podsumowanie rozważań na temat eksploracji nowych źródeł danych	188
5.5.1. Umiejętności badania nietypowych źródeł danych dla dowolnego projektu	188
Podsumowanie	189
6. Dane kateryoryczne	190
6.1. Praca z danymi kateryorycznymi	191
6.1.1. Metody obsługi danych kateryorycznych	193
6.1.2. Praca z danymi z ankiet	195
6.2. Projekt 5. Analiza ankiety badającej nastawienie programistów do narzędzi AI	196
6.2.1. Założenia projektu	196
6.2.2. Słownik danych	197
6.2.3. Oczekiwane rezultaty	198
6.2.4. Wymagane narzędzia	198
6.3. Zastosowanie metody zorientowanej na wyniki do analizy ankiety dla programistów	199

6.4. Przykładowe rozwiązanie: Jak programiści wykorzystują sztuczną inteligencję?	201
6.4.1. Analiza danych kategorycznych	201
6.4.2. Analiza danych kategorycznych z ankiet	208
6.4.3. Dotychczasowy postęp projektu	233
Podsumowanie	234
 7. Dane kategoryczne — metody zaawansowane	235
7.1. Powrót do Projektu 5. Analiza ankiety badającej nastawienie programistów do narzędzi AI	236
7.1.1. Słownik danych	236
7.1.2. Oczekiwane rezultaty	236
7.1.3. Podsumowanie dotychczasowych prac nad projektem	237
7.2. Zaawansowane metody analizy danych kategorycznych	238
7.2.1. Grupowanie wartości ciągłych w kategorie dyskretne	241
7.2.2. Stosowanie testów statystycznych do danych kategorycznych	245
7.2.3. Odpowiadanie na nowe pytanie od początku do końca	248
7.2.4. Wyniki projektu	251
7.3. Podsumowanie na temat danych kategorycznych	254
7.3.1. Umiejętności potrzebne do pracy z danymi kategorycznymi w dowolnym projekcie	254
Podsumowanie	255
 8. Szeregi czasowe — przygotowanie danych	256
8.1. Praca z szeregami czasowymi	257
8.1.1. Ukryta głębia szeregów czasowych	257
8.1.2. Jak pracować z szeregami czasowymi?	258
8.2. Projekt 6. Analiza szeregów czasowych w celu poprawy infrastruktury rowerowej	260
8.2.1. Problem	260
8.2.2. Słownik danych	261
8.2.3. Oczekiwane rezultaty	262
8.2.4. Wymagane narzędzia	263
8.3. Zastosowanie metody zorientowanej na wyniki do analizy danych o ruchu drogowym	263
8.4. Przykładowe rozwiązanie: Gdzie należy skupić się na ulepszeniach infrastruktury rowerowej?	265

8.4.1. Badanie dostępnych danych i wyodrębnianie szeregów czasowych	266
8.4.2. Dotychczasowy postęp projektu	288
Podsumowanie	290
9. Szeregi czasowe – analiza danych	291
9.1. Powrót do projektu 6. Analiza szeregów czasowych w celu poprawy infrastruktury rowerowej	292
9.1.1. Opis problemu	292
9.1.2. Słownik danych	292
9.1.3. Oczekiwane rezultaty	294
9.2. Gdzie należy skupić się na poprawie infrastruktury rowerowej?	294
9.2.1. Analiza szeregów czasowych	295
9.2.2. Wnioski i zalecenia wynikłe z projektu	334
9.3. Podsumowanie: szeregi czasowe	334
9.3.1. Umiejętności potrzebne do pracy z danymi szeregów czasowych w dowolnym projekcie	335
Podsumowanie	336
10. Szybkie prototypowanie – analiza danych	337
10.1. Proces szybkiego prototypowania	337
10.1.1. Przykład szybkiego prototypowania	339
10.2. Projekt 7. Stworzenie prototypu do analizy cen nieruchomości w Walii	341
10.2.1. Sformułowanie problemu	341
10.2.2. Słownik danych	342
10.2.3. Oczekiwane rezultaty	342
10.2.4. Wymagane narzędzia	343
10.3. Zastosowanie metody zorientowanej na wyniki w analizie danych dotyczących nieruchomości w Walii	344
10.4. Przykładowe rozwiązanie: budowa prototypu do analizy danych o cenach nieruchomości	346
10.4.1. Analiza danych przed prototypowaniem	347
10.4.2. Analiza aspektów geograficznych zbioru danych	358
10.4.3. Określanie sposobu prezentacji danych w prototypie	363
10.4.4. Dotychczasowy postęp projektu	374
Podsumowanie	374

11.	<i>Szybkie prototypowanie – tworzenie dowodu koncepcji</i>	376
11.1.	Powrót do projektu 7. Budowa prototypu do analizy cen nieruchomości w Walii	377
11.1.1.	<i>Słownik danych</i>	377
11.1.2.	<i>Oczekiwane rezultaty</i>	377
11.1.3.	<i>Podsumowanie dotychczasowych prac nad projektem</i>	378
11.2.	Tworzenie dowodu koncepcji	379
11.2.1.	<i>Przygotowanie do budowy dowodu koncepcji</i>	381
11.2.2.	<i>Wykorzystanie biblioteki streamlit do stworzenia dowodu koncepcji</i>	389
11.2.3.	<i>Wyniki projektu i dalsze działania</i>	394
11.3.	Podsumowanie na temat szybkiego prototypowania pomysłów	398
11.3.1.	<i>Umiejętności szybkiego prototypowania w dowolnym projekcie</i>	398
	Podsumowanie	399
12.	<i>Rozbudowa rozwiązania innego analityka – przygotowanie danych</i>	400
12.1.	Wyszukiwanie podobnych obiektów	401
12.2.	Kontynuowanie pracy rozpoczętej przez kogoś innego	402
12.3.	Projekt 8. Identyfikacja segmentów klientów na podstawie aktywności mobilnej	403
12.3.1.	<i>Opis problemu</i>	404
12.3.2.	<i>Słownik danych</i>	405
12.3.3.	<i>Pożądane rezultaty</i>	405
12.3.4.	<i>Wymagane narzędzia</i>	405
12.4.	Zastosowanie metody zorientowanej na wyniki do stworzenia drugiej iteracji segmentacji klientów	406
12.5.	Przykładowe rozwiązanie: tworzenie segmentów klientów	408
12.5.1.	<i>Odtwarzanie analizy innej osoby</i>	408
12.5.2.	<i>Analiza danych o zdarzeniach w celu poznania zachowań klientów</i>	421
12.5.3.	<i>Dotychczasowy postęp projektu</i>	442
	Podsumowanie	444

13.	<i>Rozbudowa rozwiązania innego analityka — segmentacja użytkowników</i>	445
13.1.	Powrót do projektu 8. Identyfikacja segmentów klientów na podstawie aktywności mobilnej	446
13.1.1.	<i>Słownik danych</i>	446
13.1.2.	<i>Oczekiwane wyniki</i>	447
13.1.3.	<i>Podsumowanie dotychczasowych prac nad projektem</i>	447
13.1.4.	<i>Segmentacja użytkowników mobilnych z użyciem grupowania</i>	449
13.1.5.	<i>Wnioski i dalsze działania</i>	459
13.2.	Podsumowanie: segmentacja i grupowanie	462
13.2.1.	<i>Podsumowanie umiejętności przydatnych w każdym projekcie</i>	462
	Podsumowanie	463
Dodatek	<i>Instrukcje instalacji Pythona</i>	465
A.1.	Instalacja Pythona	466
A.2.	Instalacja poetry	466
A.3.	Tworzenie środowiska wirtualnego	466

Łączenie teorii z praktyką w nauce o danych



W tym rozdziale

- Podejście do analizy danych oparte na rezultatach
- Kluczowe koncepcje nauki o danych na przykładzie rzeczywistych projektów
- Praktyczne rozwiązania w analizie danych i zdobywanie nowych umiejętności

Czy ten scenariusz brzmi znajomo? Właśnie otrzymałeś prośbę o dane od jednego z działów w Twojej organizacji i nie masz pojęcia, jak się do tego zabrać, a może nawet nie wiesz dokładnie, o co Cię poproszono. Po doświadczeniu z ustrukturyzowanymi danymi ze szkolenia dla początkujących rzeczywistość okazuje się chaotyczna i niepewna. Być może zastanawiasz się:

- Jak wykorzystać swoje obecne umiejętności, aby realizować projekty dla wymagających interesariuszy?
- Jak kontynuować naukę, gdy nie ma już ustrukturyzowanego środowiska szkoleniowego?
- Jak zastosować swoje ogólne umiejętności do zadań szczególnych dla danej dziedziny?
- Czego powinienesz się nauczyć w następnej kolejności?

Każdy doświadczony badacz danych powie Ci, że odpowiedzią na wszystkie te pytania jest „doświadczenie”. Realizując osiem projektów opisanych w tej książce, przyspieszysz proces zdobywania doświadczenia niezbędnego do tego, by odnieść sukces jako analityk danych.

Udoskonalisz swoje obecne umiejętności i nauczysz się nowych, wykonując projekty podobne do tych, z którymi możesz się spotkać w rzeczywistym środowisku pracy. W trakcie realizacji projektów będziesz wykonywać większość pracy samodzielnie, ale otrzymasz wskazówki, które pomogą Ci ruszyć z miejsca. Oprócz konkretnych umiejętności przedstawiam również podejście, które uczyni Cię lepszym analitykiem, skupiając się na praktycznych rezultatach. Po drodze będziesz podążać według procesu, który ułatwi Ci skuteczne nabywanie nowych umiejętności i wyciąganie maksymalnej wartości z każdego nowego wyzwania.

1.1. Zestaw narzędzi analityka danych

Istnieją pewne podstawowe umiejętności, które analitycy zdobywają na początku swojej drogi zawodowej i wykorzystują w swojej pracy. Należą do nich:

- wczytywanie zbiorów danych z różnych źródeł;
- łączenie zbiorów danych;
- manipulowanie kolumnami przez ich tworzenie, usuwanie, przekształcanie i zmienianie ich nazw;
- przeprowadzanie podstawowych analiz statystycznych, takich jak obliczanie średnich;
- eksploracja danych za pomocą wizualizacji, takich jak wykresy słupkowe, liniowe czy punktowe;
- prezentowanie wyników przez tworzenie odpowiednich wizualizacji lub projektowanie pulpitów nawigacyjnych (ang. *dashboard*).

Nie ma znaczenia, które narzędzie wybierzesz do wykonywania tych zadań. Odpowiednie narzędzia to między innymi:

- Microsoft Excel (lub jego odpowiedniki);
- narzędzia do analizy biznesowej (BI, ang. *Business Intelligence*), takie jak Tableau czy Power BI;
- języki zapytań do baz danych, takie jak SQL, choć zazwyczaj nie mają możliwości wizualizacji;
- języki programowania z funkcjami analizy danych, takie jak Python czy R.

Przykładowe rozwiązania projektów są przedstawione w Pythonie, ale nacisk kładę na rozwiązywanie problemów, a nie na szczegóły języka Python.

Realizacja projektów wzbogaci Twoje podstawowe umiejętności o kompetencje przydatne w rzeczywistych zastosowaniach, takie jak modelowanie danych,

praca z danymi kategorycznymi, pozyskiwanie danych z nietypowych źródeł oraz szybkie prototypowanie. W każdym przypadku wskażę dokładnie, jakie funkcje są potrzebne do rozwiązania problemu, abyś mógł znaleźć odpowiedni sposób na ich realizację w wybranym przez siebie narzędziu. Niektóre projekty wykorzystują modele uczenia maszynowego, ale szczegółowe omówienie tego tematu wykracza poza zakres naszych projektów.

Projekty pozwolą Ci także doskonalić „metaumiejętności”, które są kluczowe w każdym projekcie analizy danych. Są to:

- umiejętność przekształcenia ogólnego, niejasnego pytania bez żargonu technicznego w takie, na które można odpowiedzieć za pomocą danych;
- ocena dostępnych danych i określenie, czy nadają się do udzielenia odpowiedzi na pytanie;
- wiedza, jak zmienić kierunek analizy i zmodyfikować pytanie analityczne, jeśli nie można na nie odpowiedzieć z użyciem dostępnych danych;
- umiejętność komunikowania wyników osobom bez wiedzy technicznej lub nawet niezainteresowanym szczegółami technicznymi.

UWAGA Projekty praktyczne skupiają się na rozwijaniu umiejętności technicznych i rozwiązywania problemów, ale dobry analityk danych potrzebuje także innych kompetencji zawodowych. Jeśli chcesz zgłębić ten temat, polecam książkę *Skazany na sukces. Kariera w Data Science* autorstwa Emily Robinson i Jacqueline Nolis (Manning, 2021).

Przyszłość analizy danych obejmuje również narzędzia sztucznej inteligencji (AI, ang. *Artificial Intelligence*), takie jak ChatGPT i inne duże modele językowe. Narzędzia AI nie zastąpią Twojej zdolności do krytycznego myślenia, do komunikowania się z interesariuszami czy do pracy w ramach ograniczeń środowiska biznesowego. Mogą jednak przyspieszyć Twoją pracę, ponieważ pomagają automatyzować bardziej mechaniczne aspekty projektu. Wskażę przypadki, w których takie narzędzie może pomóc rozwiązać część problemu. Na przykład, jeśli nie wiesz, jak odczytać dane z pliku PDF, model AI będzie w stanie powiedzieć Ci, jak uzyskać dostęp do tej funkcjonalności w preferowanym przez Ciebie narzędziu, i dostarczy fragmenty kodu do wykorzystania. Oznacza to, że możesz szybciej rozwiązać ten problem, co zwiększy Twoją produktywność.

1.2. Podejście zorientowane na wyniki

W pracy Twój sukces zależy znacznie bardziej od osiągniętych rezultatów niż od umiejętności i wiedzy, które wnosisz do zadania. Zorientowanie na wyniki oznacza, że zawsze skupiasz się na rozwiązywaniu problemów. Powinieneś jak najszybciej stworzyć wstępne rozwiązanie — coś, co lubię nazywać *minimalną satysfakcjonującą odpowiedzią*. To właśnie przedstawiasz interesariuszom i wykorzystujesz jako podstawę do dalszych iteracji. Zastosowanie tego podejścia usprawnia Twoją pracę i zapewnia natychmiastową wartość. Poświęcasz czas na naukę czegoś,

Uwaga dotycząca terminologii

Terminy „nauka o danych” (ang. *data science*) i „analiza danych” (ang. *data analysis*) stosuję zamiennie. Przede wszystkim specjaliści od danych muszą być dobrymi analitykami, dlatego do opisanía tej dziedziny wolę używać określenia „analiza” zamiast „nauka o danych”, a na określenie osób zajmujących się tą pracą — „analityk” zamiast „specjalista od danych”. Kiedy mówię „nauka o danych”, mam po prostu na myśli „proces analizowania danych”.

Często odnoszę się też do przykładów z życia. Robię to, aby podkreślić różnicę między tym, czego uczymy się w klasie, a tym, jak wygląda praca analityka w praktyce. Ta różnica nie oznacza, że zajęcia w klasie są złe; po prostu istnieje wiele przeszkód, niekoniecznie technicznych, a w sali lekcyjnej nie sposób się nauczyć, jak sobie z nimi radzić. Skupiam się na przygotowaniu młodszych analityków do ich pracy — uczę ich wszystkiego, czego nie mogli się dowiedzieć podczas formalnego szkolenia. Mówiąc „rzeczywisty świat”, mam na myśli „poza środowiskiem edukacyjnym”.

Podobnie terminem o szerokim znaczeniu jest „interesariusz”. Będę go używać w odniesieniu do Twojego szefa, wewnętrznego klienta lub zewnętrznego odbiorcy. To osoby, od których pochodzą zlecenia, w czego wyniku masz pracę do wykonania. Interesariuszem jest każdy, kto jest docelowym odbiorcą Twojej analizy i na kogo Twoja praca będzie miała bezpośredni wpływ.

co od razu będzie przydatne, a w rezultacie osiągasz wyższy zwrot z zainwestowanego czasu. Dzięki temu podejściu zorientowanemu na wyniki nie musisz być od razu ekspertem we wszystkim. Zamiast tego, realizując kolejne projekty, będziesz dodawać do swojego zestawu narzędzi nowe, konkretne umiejętności i uczyć się efektywnego podejścia do nowych zadań. Prezentowane przeze mnie podejście zorientowane na wyniki oznacza naukę przez działanie i pracę ukierunkowaną na rezultaty.

Przyjrzyjmy się temu podejściu na przykładzie klasycznego scenariusza z branży. Załóżmy, że pracujesz w salonie samochodowym i musisz odpowiedzieć na pozornie proste pytanie swojego menedżera: „Ile samochodów sprzedaliśmy wczoraj?”. W idealnym świecie to proste zapytanie biznesowe wymagałoby jedynie przefiltrowania tabeli sprzedaży, aby wyświetlić samochody sprzedane wczoraj i podać odpowiedź. W rzeczywistości musimy głębiej przemyśleć problem, poczynawszy od zdefiniowania pojęć.

Analityczna metoda rozwiązania takiego problemu zorientowana na wyniki wygląda następująco:

1. **Zrozum problem:** Obejmuje to definicje poszczególnych słów/pojęć.
2. **Zacznij od końca:** Jaka jest minimalna możliwa odpowiedź, która mogłaby zapoczątkować dalszą dyskusję?
3. **Poszukaj dodatkowych źródeł:** Mogą to być dane, ludzie lub uprawnienia dostępu potrzebne do uzyskania tego minimalnego rezultatu.
4. **Pozyskaj dane niezbędne do uzyskania minimalnej satysfakcjonującej odpowiedzi:** Czy te dane w ogóle istnieją?

5. *Wykonaj pracę, aby uzyskać minimalną satysfakcjonującą odpowiedź:* Najlepiej, aby ten etap nie trwał dłużej niż kilka dni, co pozwoli na szybkie iteracje.
6. *Przedstaw interesariuszowi minimalny satysfakcjonujący rezultat:* Może to być zarówno nieformalna rozmowa, jak i prezentacja dla szerszego grona odbiorców.
7. *W razie potrzeby pracuj dalej:* Jeśli praca jest wartościowa, interesariusze poproszą o jej kontynuowanie.

Proces ten jest zilustrowany na rysunku 1.1. Zwróć uwagę na strzałki wychodzące z kroku 7. — podkreślają one nieliniowość procesu. W zależności od wyniku kroku 6. może być konieczny powrót do wcześniejszego etapu, nawet do samego początku.

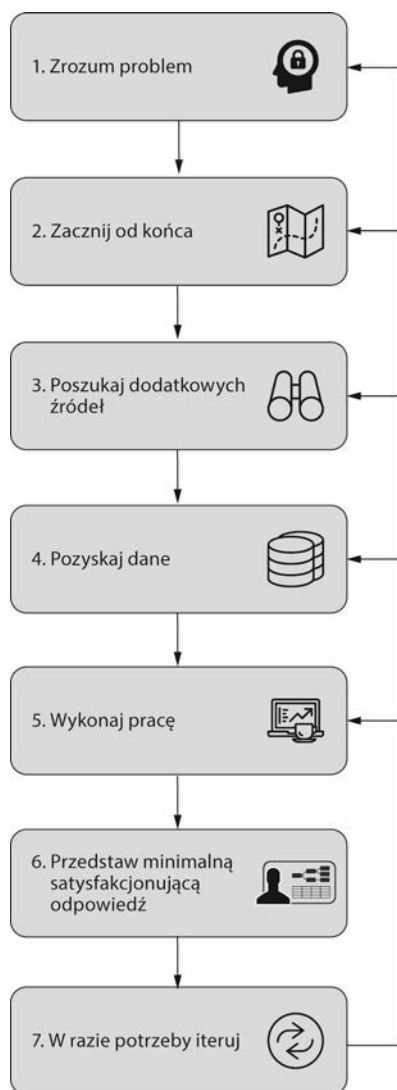
Ikony z rysunku 1.1 będą używane w całej książce przy omawianiu zastosowania tego podejścia do konkretnych projektów. Aby zastosować ten proces do naszego przykładu z samochodami, zaczynamy od kroku 1., „Zrozum problem”.

1.2.1. Zrozum problem

Co rozumiemy przez pojęcie „samochód”? Kiedy pracowałem w branży motoryzacyjnej, za każdym razem, gdy pojawiało się pytanie o pojazdy, musieliśmy doprecyzować, czy chodzi również o minibusy lub inne pojazdy, które nie są typowymi samochodami osobowymi. Czasami minibusy były istotne, a czasami nie. Następnie — co mamy na myśli, mówiąc „wczoraj”? Często z jednym wydarzeniem sprzedaży może być związanych kilka dat: dzień, w którym klient kupił produkt w sklepie, dzień wystawienia faktury, dzień zapłaty przez klienta, a nawet data rozstrzygnięcia sporu związanego ze sprzedażą. Dlatego pytając o „wczoraj”, musimy wiedzieć, którą kolumnę z datami należy wziąć pod uwagę. Nawet słowo „sprzedać” może być niejednoznaczne. Co, jeśli klient zwróci produkt później? Czy to oznacza, że odpowiedź na to pytanie z czasem się zmienia? Możemy świadomie zdecydować, że to, co wydarzyło się w przeszłości, nie jest ustalone, co skomplikuje nasze analizy. Nawiasem mówiąc, powinniśmy skodyfikować odpowiedzi na tego typu pytania i zapisać je w modelach danych specyficznych dla danej firmy. Więcej na ten temat przeczytasz w rozdziale 3.

Wynikiem tego etapu powinno być:

- jasno sformułowane pytanie, na które można odpowiedzieć za pomocą analizy;
- ustalony zakres (np. czy analiza samochodów obejmuje również minibusy?);
- jasno określona terminologia uzgodniona przez wszystkich zaangażowanych.



Rysunek 1.1. Proces podejścia zorientowanego na wyniki

Te elementy powinny być udokumentowane jako część projektu w celu zapewnienia przejrzystości i możliwości odtworzenia. Po zrozumieniu wymagań powinniśmy przejść do opracowania rozwiązania.

1.2.2. Zaczynij od końca

Zakładając, że uzgodniliśmy terminologię, możemy zacząć od końca i zdecydować, jak wygląda minimalne akceptowalne rozwiązanie. Zależy to od charakteru zapytania i tego, co nasi interesariusze chcą zrobić z informacjami. Jeśli potrzebują

szybkiego, przybliżonego oszacowania, możemy pominąć niektóre trudne pytania dotyczące typów pojazdów i dat, chyba że mamy dobry model danych, z którego możemy skorzystać. Znajomość poziomu odpowiedzi, jakiej oczekuje nasz interesariusz, jest kluczowa w określeniu, ile wysiłku należy włożyć w realizację zapytania.

Wynikiem etapu jest jasny obraz tego, jak wygląda akceptowalne rozwiązanie. Czy to dokument, prezentacja, a może działający prototyp? Po zrozumieniu tego kolejnym krokiem jest określenie wymaganych danych.

1.2.3. Poszukaj dodatkowych źródeł

Nawet w przypadku prostych zapytań ważne jest, aby zadać pytanie: „Jakie dane byłyby potrzebne do udzielenia odpowiedzi?”, a następnie: „Czy mamy te dane?”. Zaleca się, by przed podjęciem jakichkolwiek prac zbadać dostępne dane, ponieważ zdarzają się sytuacje, w których po prostu nie mamy zbioru danych bezpośrednio rejestrującego interesujące nas pojęcie. Na przykład w internetowym serwisie aukcyjnym transakcje sprzedaży mogą nie być rejestrowane bezpośrednio, ponieważ mogą odbywać się poza systemem. Musielibyśmy wiedzieć, gdzie dokonywany jest zapis, jeśli sprzedaż zostaje uzgodniona telefonicznie po zakończeniu aukcji.

Wynikiem powinien być wykaz tego, jak mogłyby wyglądać odpowiednie dane potrzebne do udzielenia odpowiedzi na pytanie. Powinno to obejmować możliwe zbiory danych już dostępne dla organizacji oraz zewnętrzne zbiory danych, które dopiero trzeba będzie pozyskać.

1.2.4. Pozyskaj dane

Teraz musimy uzyskać odpowiednie dane, albo przez ich wyodrębnienie skądś, albo przez ich stworzenie. Jeśli nasz internetowy serwis aukcyjny jest używany tylko do wystawiania ofert i licytacji, być może będziemy musieli zadowolić się wnioskowaniem o tym, kiedy coś zostało sprzedane, obserwując, od kiedy nie jest już wystawione. Jest to zbiór danych, który musielibyśmy stworzyć na podstawie surowych danych o ofertach. W takim przypadku musielibyśmy zatrzymać się na kroku 3. i zastanowić, czy tworzenie takiego zbioru danych o wywnioskowanych sprzedażach jest warte wysiłku. Będzie to zależeć od minimalnej akceptowalnej odpowiedzi, którą zdefiniowaliśmy w kroku 2.

Wynikiem powinien być konkretny zbiór danych lub jego reprezentatywny fragment, który pozwoli nam uzyskać minimalną satysfakcjonującą odpowiedź.

1.2.5. Wykonaj pracę

Krok 5. to po prostu „wykonaj pracę”. Jest to celowo ogólne i niejasne sformułowanie. Krok 5. może obejmować całość innych procesów analizy danych, z którymi się spotkałeś. Kroki 1. – 4. zapewniają, że etap wykonawczy przyniesie lepszy

zwrot z inwestycji. Na tym etapie poświęciłeś już wystarczająco dużo uwagi problemowi, więc nie rzucasz się bezmyślnie do pisania kodu, co może być pokusą, gdy masz świeże umiejętności i chęć szybkiego dostarczenia wyników. Zachęcam, żeby na kroki 1. – 4. poświęcić więcej czasu, niż pierwotnie się zakładało.

W przykładzie z samochodami na tym etapie możesz odkryć, że data sprzedaży pojazdu zmienia się w czasie w zależności od różnych zdarzeń, takich jak oczekiwanie klienta na zatwierdzenie umowy finansowej lub zgłoszenie reklamacji z powodu wcześniej niezauważonych usterek. Jako analityk nie powinieneś samodzielnie decydować, która z możliwych dat stanowi faktyczną datę sprzedaży. W takiej sytuacji należy wrócić do kroku 1. i przed kontynuowaniem analizy przeprowadzić rozmowę z interesariuszami.

Niezależnie od przebiegu analizy kluczowe jest dokumentowanie wyborów i założeń w celu zapewnienia przejrzystości i powtarzalności. W przykładowych rozwiązaniach przedstawionych dla każdego projektu udokumentowałem decyzje, które podjąłem, oraz wskazałem, gdzie moja ścieżka i ścieżka innego analityka mogłyby się rozjechać, gdyby ten podjął inne decyzje.

Wynikiem tego etapu jest to, co zostało ustalone w kroku 2., „Zaczynij od końca”. Ten artefakt powinien zostać zapisany jako część projektu i być możliwy do odtworzenia przez innych analityków. W rozdziałach 12. i 13. przedstawiam projekt, w którym właśnie to robimy.

1.2.6. Przedstaw minimalną satysfakcjonującą odpowiedź

Kluczowym elementem podejścia zorientowanego na wyniki jest prezentowanie rezultatów wcześniej, niż byśmy chcieli. Szukamy zadowalającego rozwiązania, więc powinniśmy omawiać wstępne wyniki tak wcześnie, jak to możliwe, aby umożliwić iterację. Zespoły ds. danych w przeszłości ponosiły porażki, gdy izolowały się od reszty firmy, by po miesiącach wyjść z pracą, o którą nikt nie prosił i która nie miała namacalnej wartości biznesowej.

Wynikiem tego etapu są notatki ze spotkania lub prezentacji dla naszych interesariuszy. Rezultaty powinny zostać również zapisane jako część projektu. Czy poproszono o dodatkową pracę? Jeśli tak, to jaką?

1.2.7. W razie potrzeby iteruj

Na koniec prawie zawsze będziesz musiał rozwinąć swoje pierwsze rozwiązanie. To ważny krok w każdym procesie analizy danych, ponieważ podkreśla nieodłączną niepewność tej pracy. Nigdy nie będziesz miał wszystkich odpowiedzi z góry, w tym odpowiedzi na pytanie „Ile czasu zajmie ta analiza?” — i to jest w porządku. Zaakceptuj, że pierwsza wersja będzie niezbędna do osiągnięcia satysfakcjonującego rozwiązania i że im szybciej ta pierwsza wersja powstanie, tym lepiej. Dodatkową korzyścią ze stosowania tego iteracyjnego podejścia jest to, że stajesz się

Czy ktoś korzysta z Twojej pracy?

Kiedyś na jednej konferencji uczestniczyłem w interesującej prezentacji, w której zespół od danych przedstawił skomplikowany model prognozowania sprzedaży. Prezentacja miała czysto techniczny charakter i opowiadała o tym, jak zespół doszedł do zastosowania hierarchicznego podejścia bayesowskiego. W części poświęconej pytaniom i odpowiedziom zadałem trudne pytanie: „Czy ktoś korzysta z waszych prognoz, a jeśli tak, to w jakim celu?”. Naukowcy zajmujący się danymi nieśmiało przyznali, że wciąż pracują nad przekonaniem biznesu o wartości ich prognoz. Podejrzewam, że gdyby wcześniej podjęli tę rozmowę, mogliby uniknąć wielu niepotrzebnych działań.

lepsy w szacowaniu, ile czasu zajmie Twoja praca, ponieważ musisz oszacować tylko jedną iterację, a nie cały projekt.

Stosując podejście zorientowane na wyniki, z czasem rozwinięsz intuicję dotyczącą podejmowania decyzji analitycznych, które pomogą Ci skupić się na dostarczeniu pragmatycznego rozwiązania zamiast niepotrzebnie wyczerpującej odpowiedzi już na samym początku. Oczywiście, będziesz musiał też nauczyć się obsługi nowych narzędzi. Czasami te narzędzia będą używane tylko do analizy, którą masz przed sobą. Następna analiza może wymagać innych narzędzi. To normalne — narzędzia przychodzą i odchodzą, ale wiedza i koncepcje, których się nauczysz, pozostaną takie same, a to one sprawiają, że ktoś jest dobrym analitykiem.

Ważne jest, aby zauważyć, że to podejście nie zastępuje dogłębnej analizy, gdy jest ona konieczna. Nie zachęcam nikogo, żeby się uczył, jak chodzić na skróty. Jednak kluczową umiejętnością w analityce jest zdolność określania najwyższego zwrotu z inwestycji własnego czasu, co często oznacza przyjęcie podejścia „wszerz” zamiast „w głąb”.

1.3. Struktura projektu

Projekty zostały zaprojektowane tak, aby odzwierciedlały typowe problemy, z którymi spotykają się analitycy w branży. Każdy projekt jest skonstruowany w taki sposób, aby można było go zrealizować z użyciem podstawowych umiejętności analitycznych i nie zajmował zbyt dużo czasu. Znaczenie ma tu wiele czynników: doświadczenie w korzystaniu z narzędzi, znajomość tematyki, szybkość przyswajania nowych pojęć i tak dalej. Ogólnie rzecz biorąc, pierwsze podejście do rozwiązania każdego projektu powinno zająć od dwóch godzin do dwóch dni. To szeroki zakres, ale nawet dwudniowy „sprint”, by dostarczyć coś, co można omówić z interesariuszem, uznaje się za krótki czas realizacji.

Projekty nie tylko odzwierciedlają typowe problemy branżowe, ale również zaprojektowano je tak, aby wypełniały konkretne luki w umiejętnościach między wstępnym szkoleniem a rzeczywistością zawodową. Ich przegląd jest zawarty w tabeli 1.1.

Tabela 1.1. Podsumowanie ośmiu projektów z książki

Projekt	Umiejętności analizy danych	Rozdział(y)
Analiza wydatków klientów detalicznych w różnych obszarach geograficznych	ekstrakcja ustrukturyzowanych danych z tekstu swobodnego	2.
Wyodrębnianie unikatowych rekordów klientów z transakcji w sklepie internetowym	modelowanie danych	3.
Określanie i znajdowanie najlepiej sprzedających się produktów w sklepie internetowym	definiowanie wskaźników	4.
Analiza wpływu pandemii koronawirusa na branżę filmową	pozyskiwanie danych z plików PDF	5.
Badanie postaw programistów wobec narzędzi AI na podstawie ankiety	obsługa danych kategorycznych	6., 7.
Wskazanie potencjalnych usprawnień infrastruktury rowerowej	praca z danymi szeregów czasowych	8., 9.
Budowa prototypu aplikacji do eksploracji rynku nieruchomości w Walii	szybkie prototypowanie	10., 11.
Kontynuowanie pracy innego analityka i tworzenie segmentów klientów na podstawie ich aktywności w telefonach komórkowych	iteracja na podstawie cudzej pracy, segmentacja klientów, grupowanie	12., 13.

Każdy projekt ma następującą strukturę:

- Zaczyna się od ogólnego opisu i danych. Dalej musisz radzić sobie sam. Ta część wystarczy, jeśli czujesz się pewnie w formułowaniu planu działania na podstawie dość ogólnego pytania analitycznego lub chcesz poćwiczyć tę umiejętność.
- Następnie znajdziesz bardziej szczegółowy opis możliwego podejścia do rozwiązania. Nie będzie tu fragmentów kodu, ale jeśli po ogólnym opisie nie będziesz wiedział, od czego zacząć, ta część powinna pomóc Ci w rozpoczęciu pracy nad rozwiązaniem.
- Na końcu zawsze umieszczam przykładowe rozwiązanie. Omawiam w nim możliwe sposoby podejścia do problemu, jakie założenia trzeba przyjąć, jak można zmodyfikować pytanie analityczne, biorąc pod uwagę ograniczenia danych, oraz jak powinien wyglądać satysfakcjonujący wynik. W przykładowym rozwiązaniu przyjmuję pewne konkretne założenia i decyzje, które mogą różnić się od Twoich. To normalne. Celem nie jest to, abyś doszedł do takiego samego rozwiązania jak ja — moje rozwiązanie nie jest jedynym słusznym. Wręcz przeciwnie, zwracam uwagę na miejsca, w których analiza mogła pójść inną drogą. Celem jest ćwiczenie procesu od określenia problemu do uzyskania minimalnego akceptowalnego rozwiązania.

Traktuj ten zbiór realistycznych projektów analitycznych jako towarzysza ćwiczeń, który pomoże Ci przyzwycząć się do zagłębiania się w problemy bez znajomości odpowiedzi czy nawet potrzebnych narzędzi. Ucząc się poprzez projekty praktyczne, będziesz:

- miał przed sobą konkretny cel, do którego dążysz;
- uczył się umiejętności niezbędnych do znalezienia rozwiązania, zamiast przyswajać wiedzę bez określonego celu;
- poznawał zagadnienia wszerek, a nie w głąb.

Podejście zorientowane na rezultaty przygotowuje Cię do zastosowania tych metod w dowolnym przedsięwzięciu. Takie projekty powinny inspirować Cię do dalszej analizy danych i do tego, żeby wykorzystywać schemat „rozwiązywania problemów” w nauce, jednocześnie gromadząc nowe umiejętności i budując przydatne portfolio do zaprezentowania potencjalnym pracodawcom.

Zaczynamy! Kolejnym krokiem jest podjęcie się pierwszego projektu — analizy danych demograficznych klientów dla hipotetycznego brytyjskiego sprzedawcy detalicznego. Mam nadzieję, że spodoba Ci się ta podróż!

Podsumowanie

- Spodziewaj się różnicy między formalnym wykształceniem w dziedzinie nauki o danych a rzeczywistością zawodową.
- Pragmatyczne podejście i skoncentrowanie się na wynikach pomagają poruszać się w niepewności, jaka nieodłącznie towarzyszy analizie danych.
- Ćwiczenie tego podejścia zorientowanego na rezultaty na podstawie przykładów rzeczywistych danych pomoże Ci:
 - skupić się na rozwiązywaniu problemów;
 - lepiej identyfikować dodatkowe aspekty;
 - rozwinąć intuicję dotyczącą podejmowania decyzji analitycznych, aby szybciej dostarczać pragmatyczne rozwiązania problemów interesariuszy;
 - stworzyć portfolio projektów opartych na rzeczywistych danych, które wyróżni Cię jako kandydata.

PROGRAM PARTNERSKI

— GRUPY HELION —



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW w działający bankomat!

Dowiedz się więcej i dołącz już dzisiaj!

<http://program-partnerski.helion.pl>

GRUPA
Helion

Analiza trudnych danych w rzeczywistym świecie? Cóż... to skomplikowane. Trzeba sobie poradzić z takimi wyzwaniami jak niewiarygodne źródła danych, niejednoznaczne zapytania i niekompatybilne formaty. Rzeczywistość wykracza poza czyste, ustrukturyzowane przykłady z podręczników. Jak więc z trudnymi danymi radzą sobie zawodowcy?

Dobrze napisana książka, pokazująca podejścia do rozwiązywania wielu rodzajów problemów w analizie danych.

— Naomi Ceder, członkini Python Software Foundation

Za sprawą tej książki poznasz metodologię rozwiązywania różnorodnych problemów związanych z pracą na danych. Poszczególne techniki zostały tu wyjaśnione krok po kroku, tak by odzwierciedlić tok myślenia i sposób działania profesjonalnego analityka. Znajdziesz tu osiem często spotykanych scenariuszy, z którymi można się zetknąć w tej dziedzinie. Stosując zorientowane na wyniki podejście autora, nauczysz się modelować dane poprzez odnajdywanie rekordów klientów, pracę ze słabo zdefiniowanymi wskaźnikami, wydobywanie informacji z plików PDF i wiele więcej! Dowiesz się również, jak przejmować niedokończone projekty i tworzyć szybkie prototypy z rzeczywistymi danymi.

Możesz podglądać pracę wytrawnego analityka danych, który wyjaśnia, jak radzić sobie ze wszystkimi nieprzyjemnymi pułapkami!

— Randy Au, Google

W książce:

- rozkładanie problemów na części pierwsze
- praca z niejasnymi wskaźnikami
- modelowanie danych
- manipulacja danymi kategorycznymi

Jasna i powtarzalna metodologia ułatwi Ci pewne i szybkie dochodzenie do odpowiedzi!

— Shaun McGirr, DevOn Software Services

David Asboth jest doświadczonym analitykiem danych, łączącym wiedzę z zakresu tworzenia oprogramowania i projektowania systemów. Współprowadzi podcast *Half Stack Data Science*, występuje także na konferencjach, między innymi na londyńskim Data Science Festival.

Doskonałe źródło dla każdego aspirującego analityka danych!

— Andrew R. Freed, IBM

Helion 	KOD KORZYŚCI Siegnij po więcej! ▶ 
 helion.pl	ISBN 978-83-289-3308-8
 HELION S.A. ul. Kościuszki 1c 44-100 Gliwice tel.: 32 230 98 63 helion@helion.pl	 9 788328 933088
Cena: 119,00 zł	