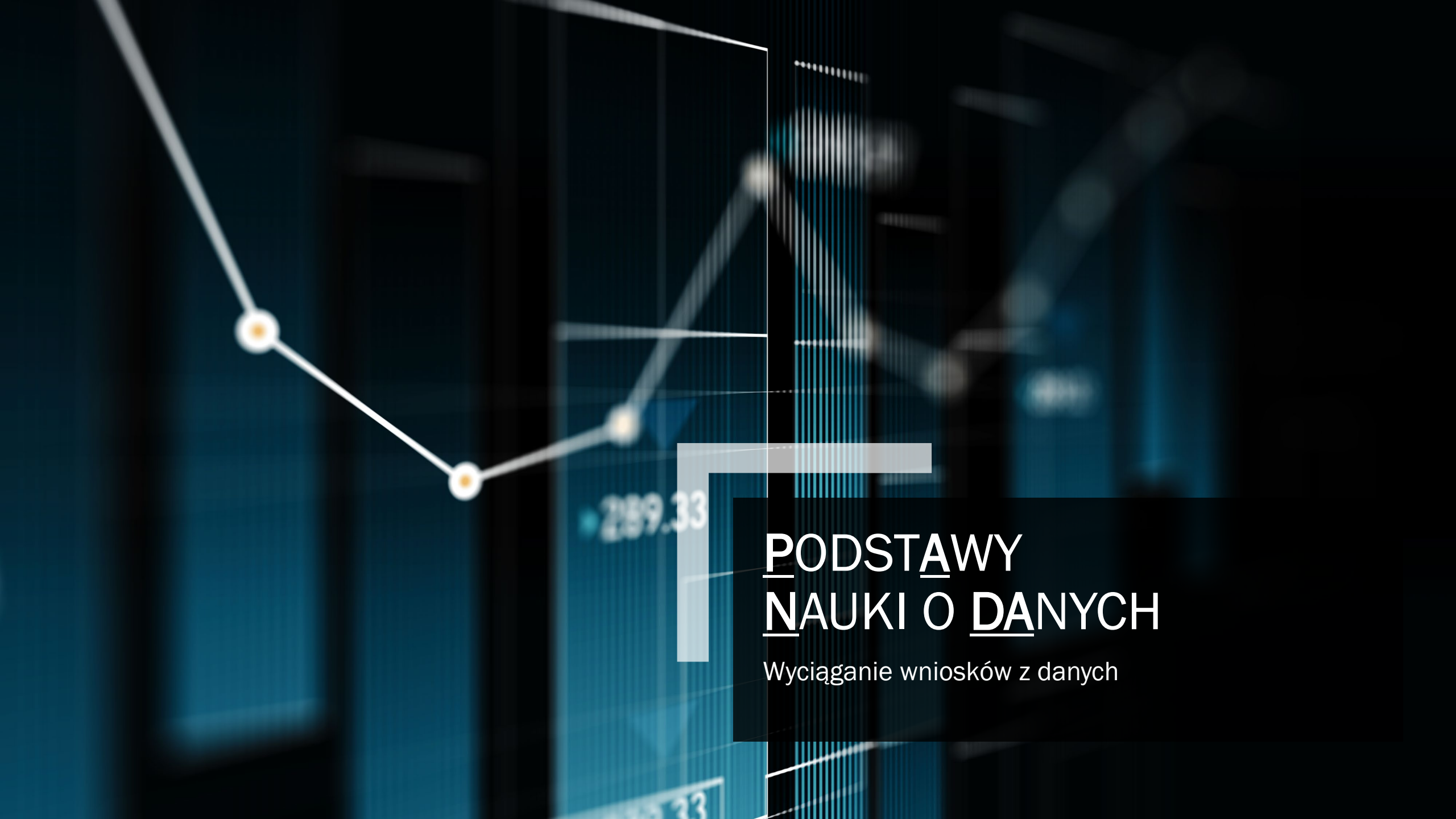
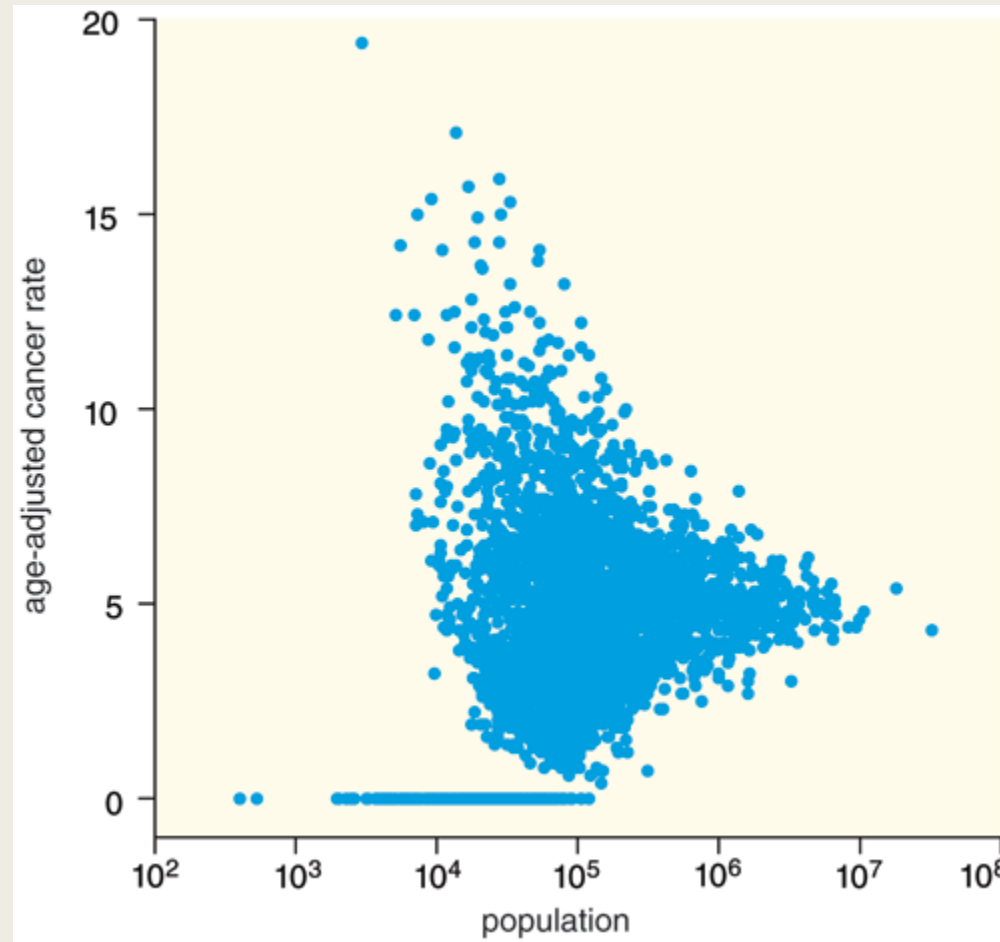




PODSTAWY NAUKI O DANYCH

Wyciąganie wniosków z danych



Howard Weiner, The Most Dangerous Equation, *American Scientist*, vol. 95(3), 2007



99 %



?

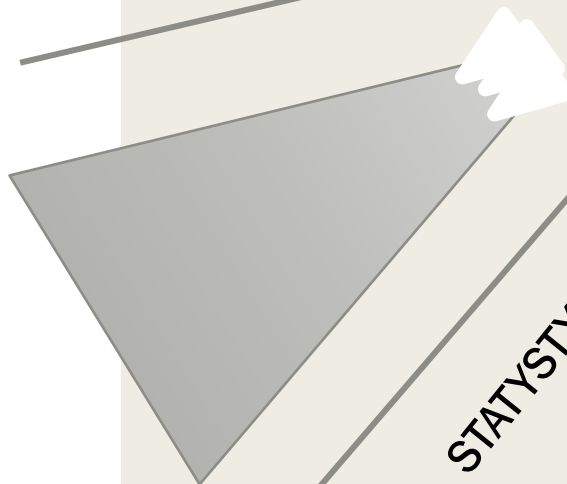


Czy moje **decyzje** bardziej opierają się na **danych** czy na **wiedzy**?



populacja

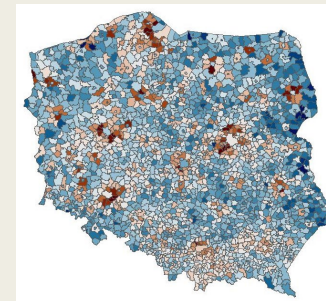
PRAWDOPODOBIENSTWO



STATYSTYKA

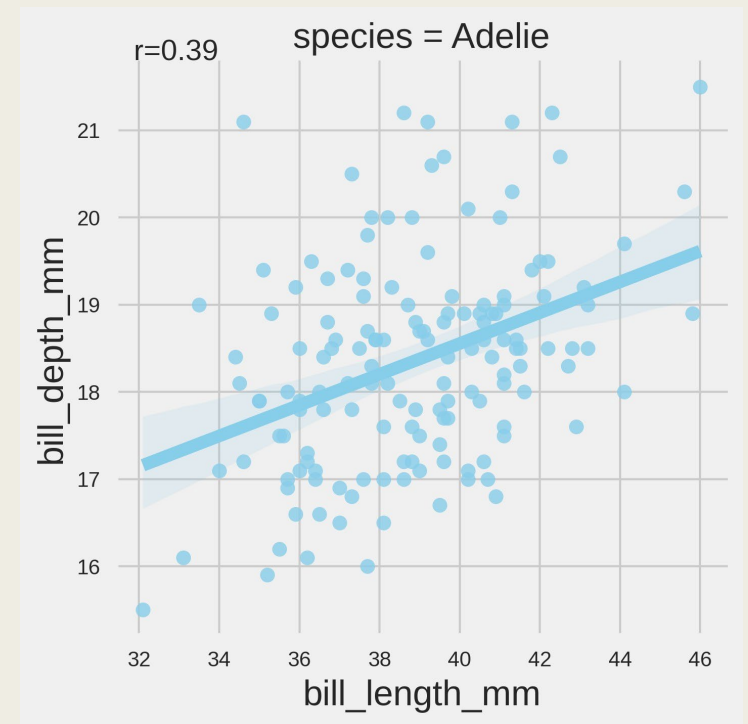
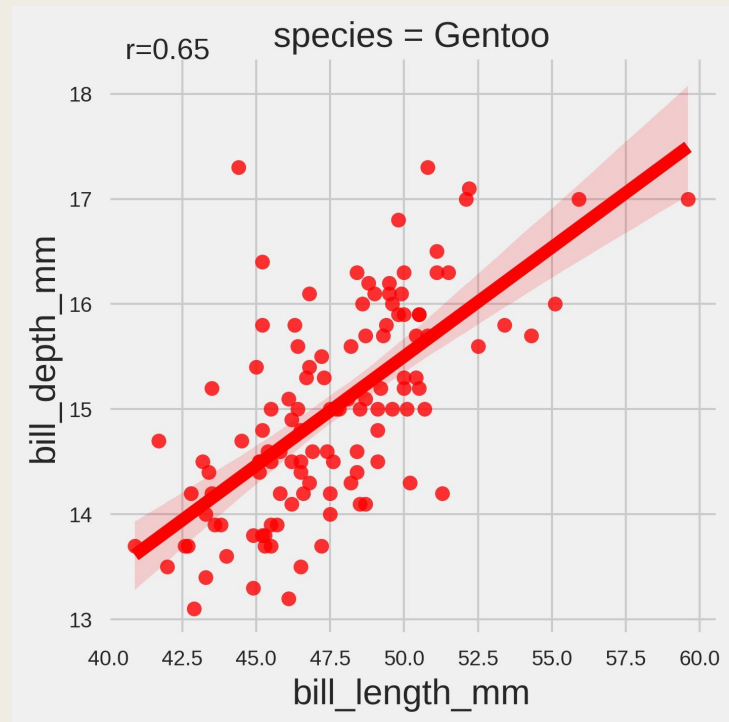
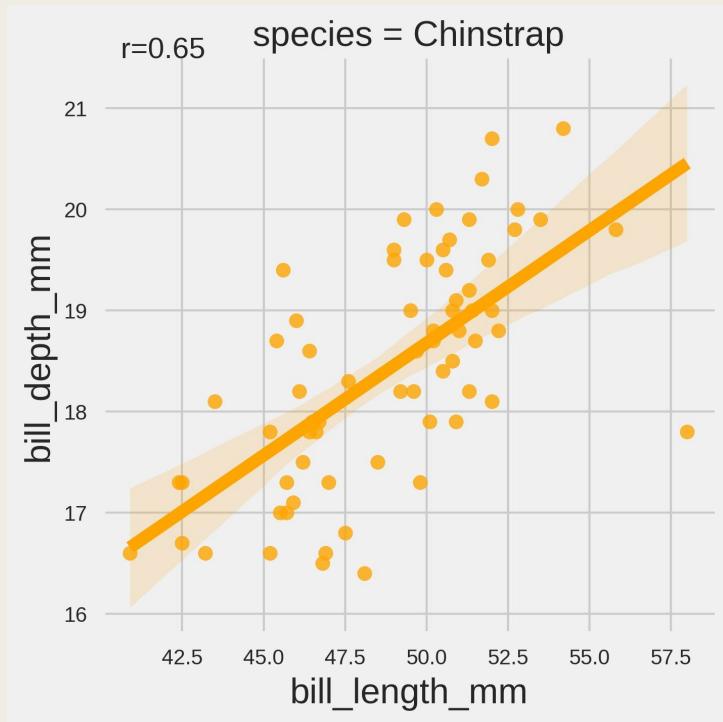


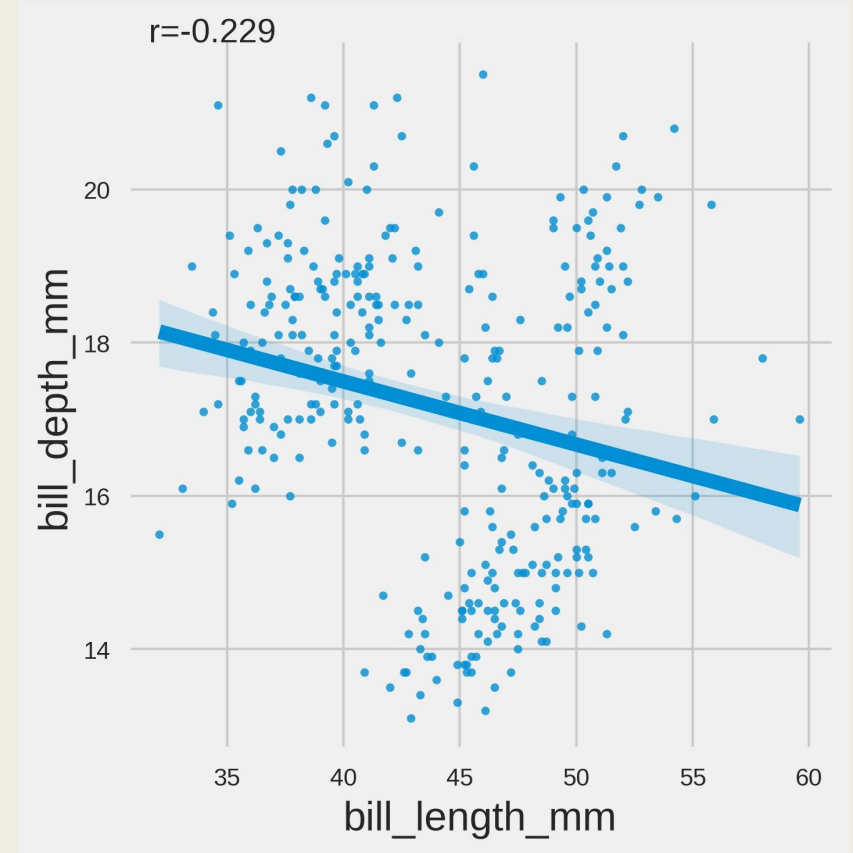
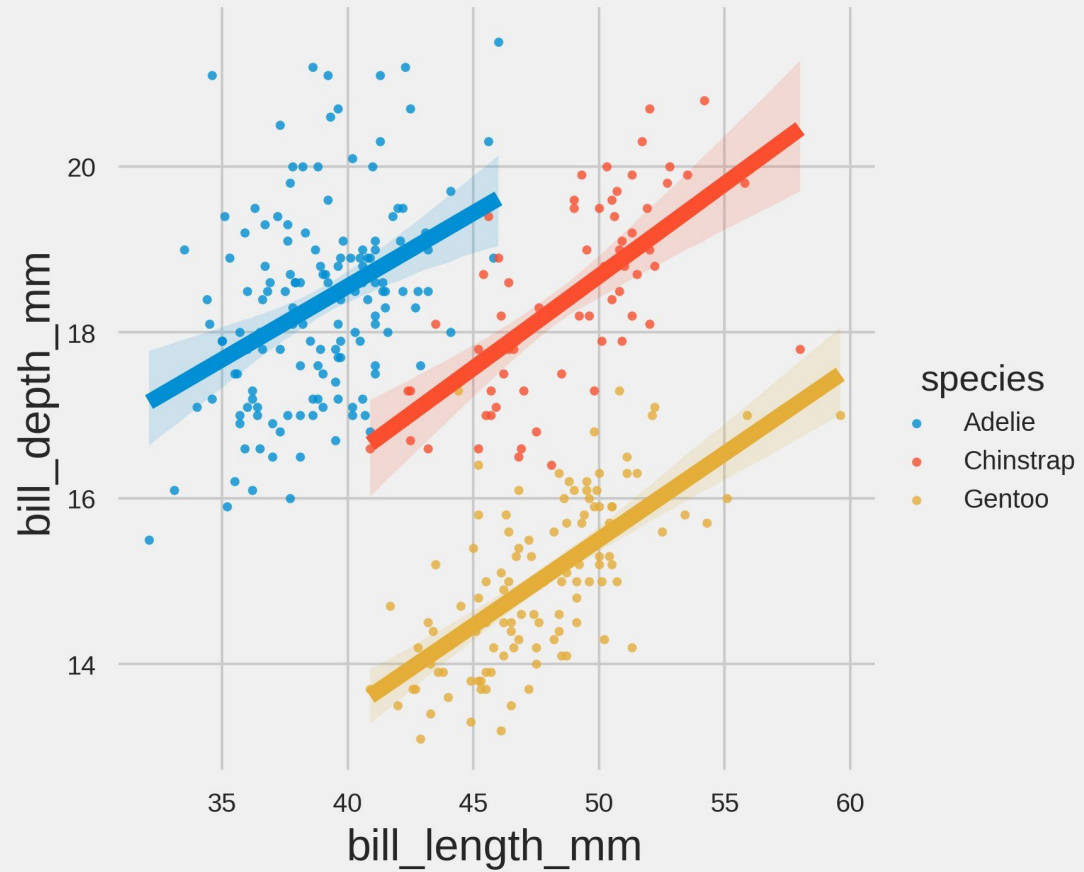
próba



2019 North Korean Supreme People's Assembly Election



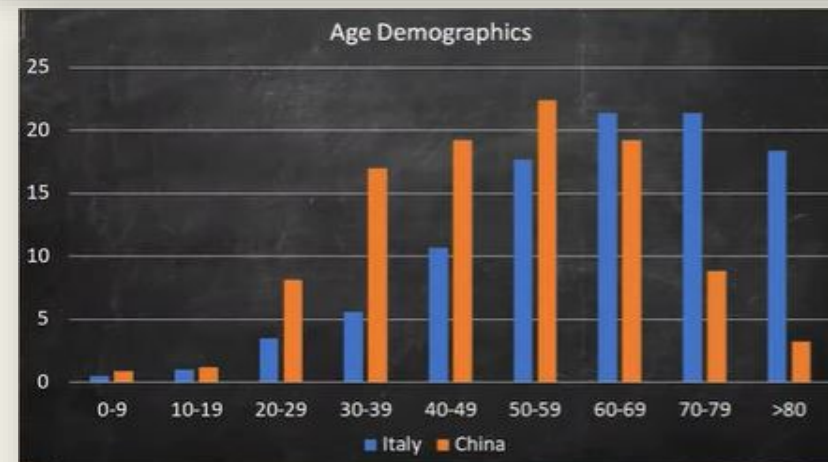
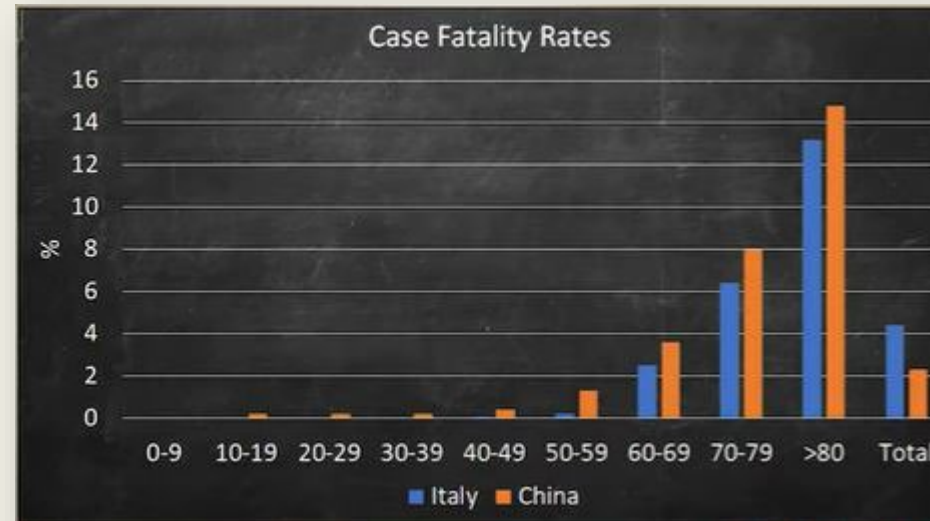




von Kügelgen, J., Gresele, L., & Schölkopf, B. (2021).

Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects.

IEEE Transactions on Artificial Intelligence, 2(1), 18-27.





Niepalący	Wyleczonych	Niewyleczonych	Razem
Nowy lek	85	15	100
Sławy lek	300	100	400
	385	115	500

$$\frac{85}{100} > \frac{300}{400}$$

więc nowy lek jest lepszy



Palący	Wyleczonych	Niewyleczonych	Razem
Nowy lek	40	160	200
Sławy lek	30	170	200
	70	330	400

$$\frac{40}{200} > \frac{30}{200}$$

więc nowy lek jest lepszy



+



Ogółem	Wyleczonych	Niewyleczonych	Razem
Nowy lek	125	175	300
Sławy lek	330	270	600
	455	445	900

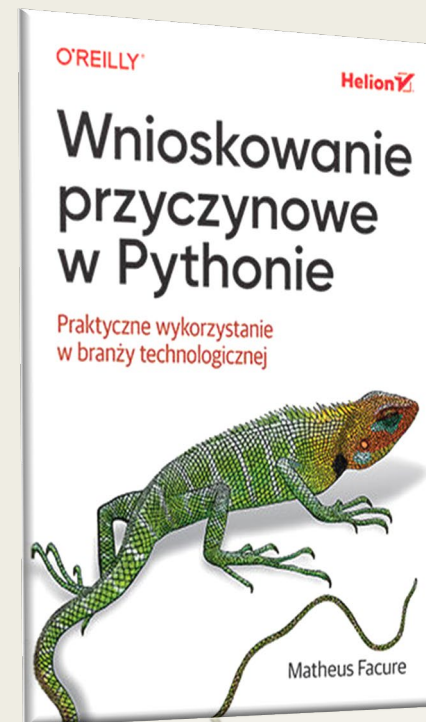
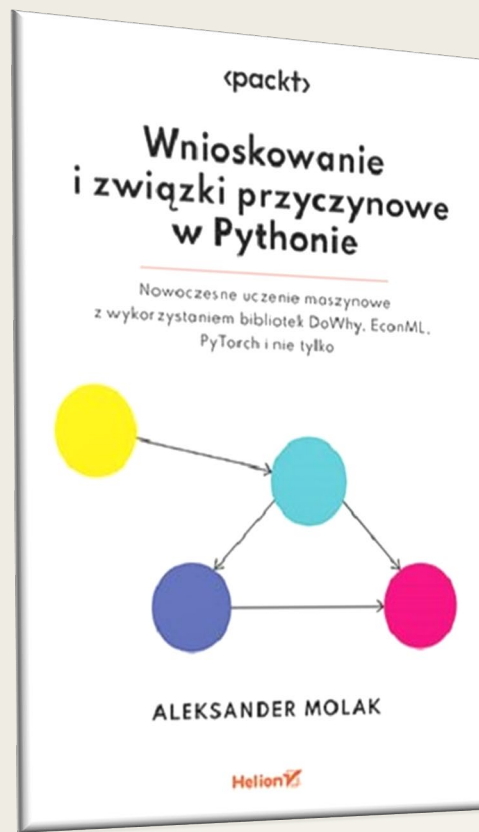
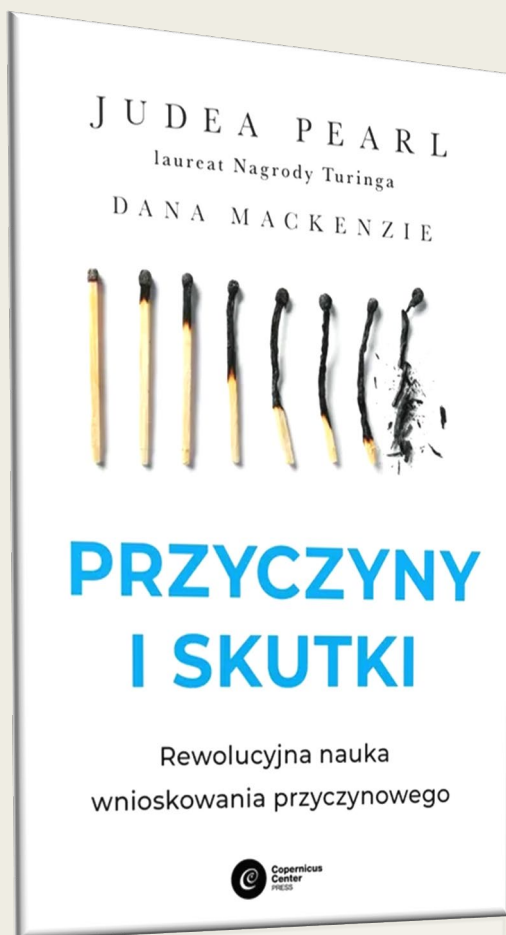
$$\frac{125}{300} < \frac{330}{600}$$

więc sławy lek jest lepszy

Miej ograniczone zaufanie do danych

- Czy dane związane są z celem modelowania? (*Big Data*)
- Czy dane faktycznie mierzą to co mierzą? (*poczucie szczęścia, palenie papierosów*)
- Jakich danych nie chciano podać i jak to zakodować? (*zarobki, wiek*)
- Co się stało z danymi nietypowymi? (*oskubywanie danych*)
- Jakich danych nie widzę? (*błąd przeżywalności*)
- Czy kryterium doboru do próby odpowiada celom modelowania? (*tendencyjna próba*)
- Czy więcej danych to lepiej? (*tendencyjna próba*)
- **Czy rozumiem proces generujący dane?**

Czy rozumiem proces generujący dane?



koniec