CZYTANIE DANYCH I STATYSTYKI

EKSPLORACJA DANYCH

Eksploracja danych

1. Struktura danych - zmienne, obserwacje, skale pomiaru
2. Dane istotne, które zignorować, na które uważać
3. Obserwacje odstające - wykrywanie i radzenie sobie z nimi
4. Jakość danych - jak ocenić, skąd wiemy że statystyki nas nie oszukują?

Struktura danych

Dane składają się z obserwacji i zmiennych

Obserwacja to jeden wiersz w danych. Np. informacje o jednym z klientów

Zmienna to jedna z kolumn. Np. wiek lub zarobki

Struktura danych

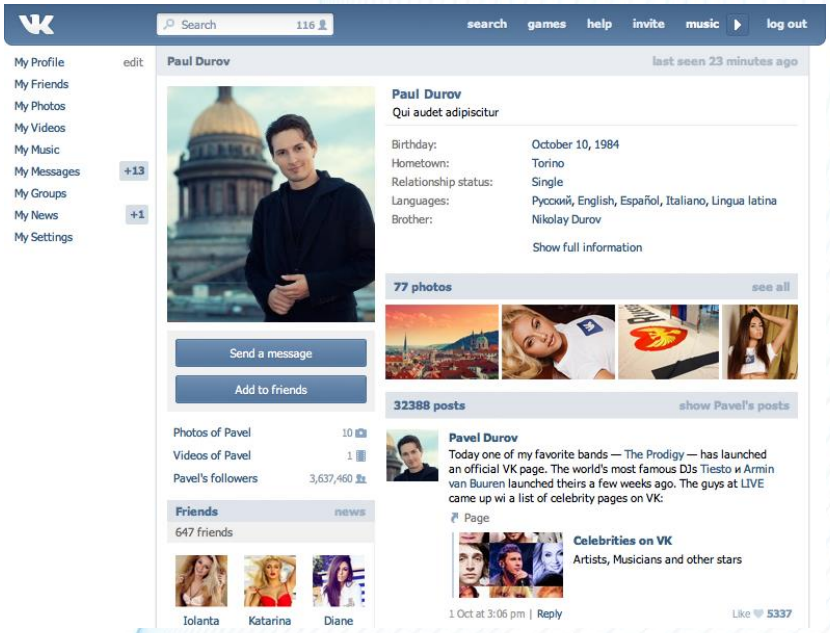
- Dane mogą być przedstawione w postaci tabelarycznej. Z takimi danymi mamy do czynienia w Excelu i bazach danych SQL

MONTH	DEMAND	CAPACITY
APR	46 193	29 263
MAY	49 131	28 037
JUN	50 124	21 596
JUL	48 850	25 895
AUG	47 602	25 813
SEP	43 697	22 427
OCT	41 058	23 605
NOV	37 364	24 263
DEC	34 364	24 243

Struktura danych

- Dane mogą być również przedstawione w postaci nietabelarycznej. Z takimi danymi mamy do czynienia w plikach JSON i bazach NO SQL

- Przykładem mogą być na przykład dane na profilach społecznościowych



Skale danych

- **Skala nominalna** – wartości w tej skali nie są uporządkowane według określonego porządku, a jedyną relacją porównującą dwie wartości jest równość. Jedną z podgrup takiej skali są skale dychotomiczne, które przyjmują wyłącznie dwie wartości, np. odpowiedź na pytanie „tak” lub „nie”.

Skale danych

- **Skala porządkowa** – w tym przypadku wartości są określone w pewnym porządku, np. ze względu na jakąś właściwość. Poszczególne przypadki można uporządkować według wskazanego kryterium, natomiast nie dostarcza ona żadnych innych danych. Dla przykładu wskazując wykonywany zawód jako „manager” lub „pracownik biurowy” wiadomo, że pierwszy z nich jest stanowiskiem wyższym, jednak nie wiadomo o ile.

- Przykładem takiej skali jest skala Likerta

Skale danych

- **Skala Likerta**

1. – zdecydowanie nie zgadzam się,
2. – raczej się nie zgadzam,
3. – nie mam zdania,
4. – raczej się zgadzam,
5. – zdecydowanie się zgadzam

Skale danych

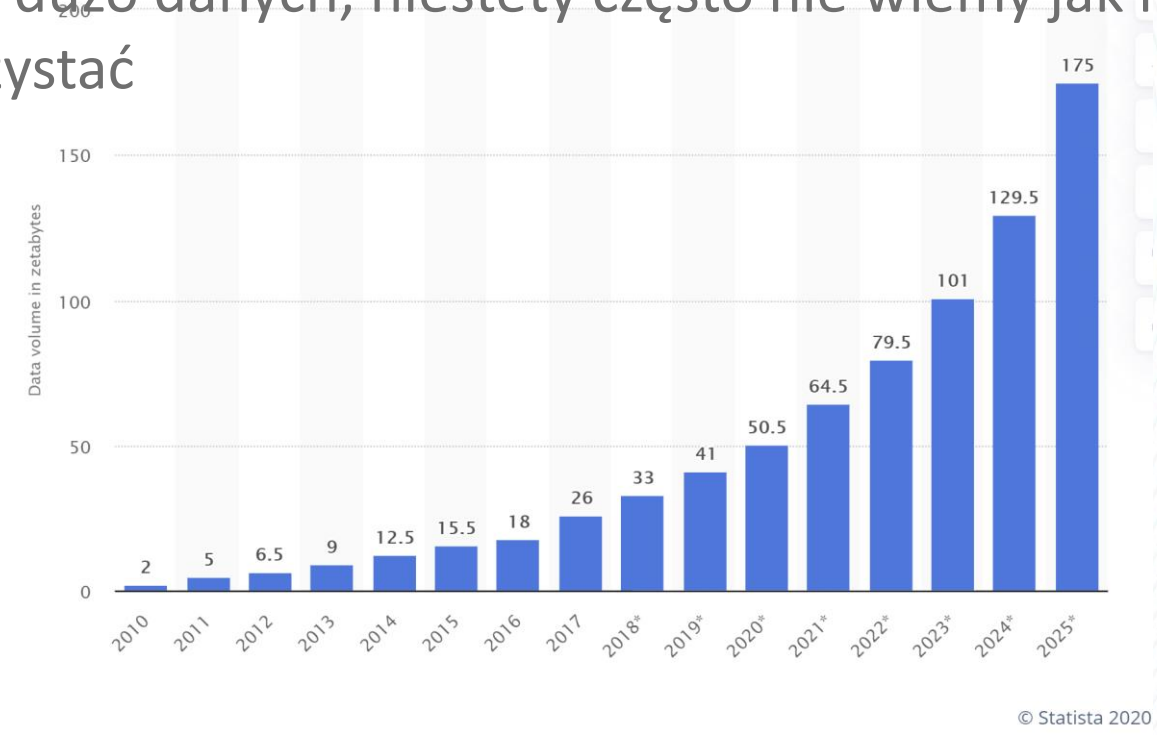
- **Skala przedziałowa**, inaczej zwana interwałową – w tym przypadku różnice między dwoma zmiennymi można obliczyć i mają one interpretację w świecie rzeczywistym, jednak dzielenie tych wartości przez siebie nie ma sensu, gdyż nie daje żadnego wymiernego wyniku. Przykładami zmiennych przedziałowych są daty, stopnie Celsjusza.

Skale danych

- **Skala ilorazowa** – stosunki między zmiennymi maja interpretację w świecie rzeczywistym. Możliwe jest dzielenie zmiennych przez siebie, a ponadto skala ta posiada cechy poprzednich trzech skal. Jej cechą charakterystyczną jest posiadanie punktu zero, oznaczającego brak danej cechy. Skala ta pozwala na uzyskanie dokładnych wartościowo różnic pomiędzy badanymi cechami statystycznymi.

Dane istotne i nieistotne

- Zbieramy dużo danych, niestety często nie wiemy jak możemy je wykorzystać



Dane zbierane każdego roku na świecie w latach 2010-2017 i prognoza.

Dane istotne i nieistotne

- Dane, które posiadamy mogą być nieprzydatne do naszego celu

- Jak możemy to zbadać?
  - Wiedza ekspercka
  - Współczynnik korelacji
  - Wizualizacja danych na wykresach
  - Budowa modeli statystycznych

Dane istotne i nieistotne

- Dane mogą być dla nas istotne i potrzebne ale błędne, zebrane w nieodpowiedni sposób

- Niestety nasza analiza będzie tylko tak dobra jak dobre sa nasze dane. Jeśli nasze dane są błędnie, nie wyciągniemy z nich prawidłowych wniosków

Obserwacje odstające

Inaczej **outliery** to obserwacje, które są relatywnie daleko od pozostałych.

Skąd biorą się outliery?

Osoba	Zarobki
Jan Kowalski	100k
Anna Nowak	75k
Jan Iks	99k
Aneta Zwyczajna	60k
Bill Gates	**99999999k**

Obserwacje odstające

- Te dane mogą być prawidłowe ale mogą też być wynikiem błędu lub oszustwa
- Niezależnie od powodu ich pochodzenia, będą zaburzały nasze statystyki, więc należy rozważyć ich usunięcie
- Przyjmuje się że outlierem jest obserwacja, która jest dalej niż 3 odchylenia standardowe od średniej

Są również outliery, które nie mają bardzo wysokich lub niskich wartości. Jak to jest możliwe?

Obserwacje brakujące

- Czasami trafiamy na braki w danych.
- Zakodowane są typowo jako puste miejsce, NA, NaN lub „-”

Co może być powodem braków w danych?

Prague	34
Warsaw	12
London	NA
Berlin	98

Obserwacje brakujące

Braki w danych możemy podzielić na takie, które są brakujące:

- Sposób losowy
- Za którymi kryją się pewne wzorce

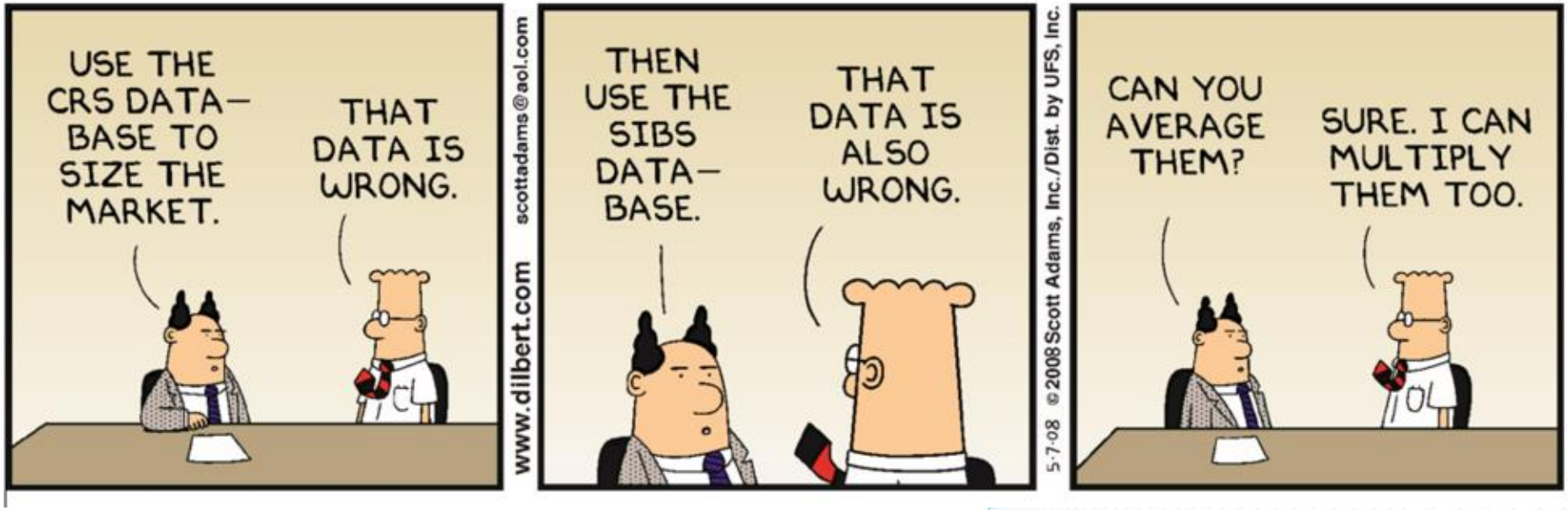
Czy potrafisz podać przykład dwóch typów braków?

Obserwacje brakujące

W zależności od tego co jest przyczyną braków inaczej będziemy musieli podejść do analizy takich danych.

Osoba	Czy zażywa Pan/Pani narkotyki?
A	NIE
B	NIE
C	NA
D	NIE
E	NIE

Złe dane – złe decyzje!



Jakość danych

- Analiza jest tylko tak dobra jak dobre są dane. Innymi słowy, analiza wykonana na błędnych danych będzie błędna
- Co to znaczy, że dane są złej jakości? Kiedy mamy problem?
- Jak możemy zapobiegać problemom w jakości danych?
- Co jeśli już wiemy, że nasze dane są złej jakości?

Ćwiczenie

1. Otwórz arkusz Eksploracja danych
2. Obejrzyj znajdujące się w nim dane
3. Sprawdź czy nie ma braków w danych
4. Sprawdź czy nie ma błędnych danych lub obserwacji odstających
5. Chcesz wyszukać w danych „podejrzane” samochody. Powypadkowe, uszkodzone z przekręconym licznikiem… które zmienne wydają ci się istotne do tego celu, a które można odrzucić?