CZYTANIE DANYCH I STATYSTYKI

PREZENTACJA I INTERPRETACJA DANYCH

Prezentacja i interpretacja danych

1. Rodzaje zmiennych i dlaczego to ma znaczenie dla prawidłowego przedstawiania danych
2. Różne metody prezentacji wyników, liczby da się pokazać nie tylko za pomocą tabelek
3. Jak szukać zależności między zmiennymi?
4. Najczęstsze błędy w prezentacji i interpretacji danych i jak ich uniknąć?

Rodzaje zmiennych

Zmienne możemy podzielić na dwie kategorie:

- Zmienne ilościowe
- Zmienne jakościowe

Liczbowe ciągłe

- W skali ilorazowej lub przedziałowej
- Przedstawione za pomocą liczb rzeczywistych

SalePrice
208500,78
181500,00
223500,08
140000,54
250000,12
143000,65

Liczbowe dyskretne

- W skali zliczeniowej, ilorazowej lub przedziałowej
- Przyjmują tylko wartości całkowitoliczbowe

YearBuilt
2003
1976
2001
1915
2000
1993

Binarne (logiczne)

Fence
NO
NO
YES
NO
YES
NO
YES

- Te zmienne przyjmują tylko dwie wartości
- Często jest to 1-0
- Lub PRAWDA-FAŁSZ

Kategorie nominalne

Neighborhood
CollgCr
Veenker
CollgCr
Crawfor
NoRidge
Mitchel

- Te zmienne przyjmują wiele wartości
- Mogą być przedstawione za pomocą liczb (na przykład numer kategorii) lub za pomocą tekstu

Kategorie uporządkowane

Overall Quality
7
6
7
7
8
5

- Kategorie, które można posortować większy/mniejszy, ale na których nie da się wykonywać obliczeń
- Mogą być przedstawione za pomocą liczb (na przykład klasa jakości) lub za pomocą tekstu

Zmienne tekstowe

Comment
Everything ok
I was very satisfied!
Thumbs up!
Big disapointment…
OK

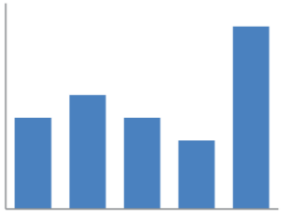
- Dane w postaci tekstu
- Każda obserwacja ma unikatową wartość
- Takich danych nie da się łatwo agregować

Prezentacja wyników

91%

Simple text

Scatterplot



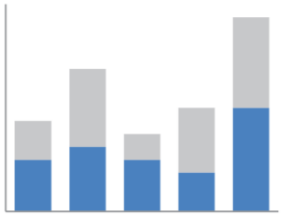
Vertical bar

Horizontal bar

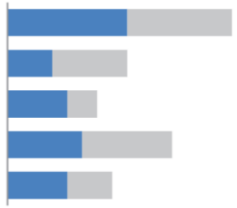
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Table

Line



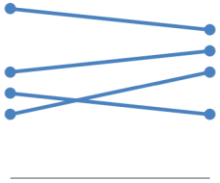
Stacked vertical bar



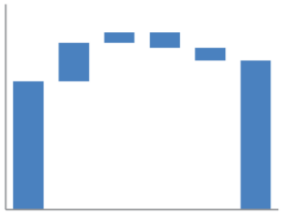
Stacked horizontal bar

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

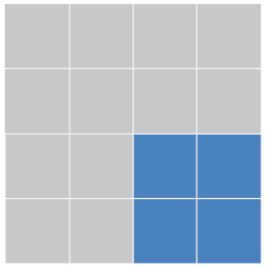
Heatmap



Slopegraph



Waterfall



Square area

Zależności między zmiennymi

Sposoby badania zależności:

- Współczynnik korelacji
- Tabela krzyżowa
- Wykres typu scatterplot
- Wykres słupkowy lub boxplot, gdy jedna ze zmiennych to kategorie

Ćwiczenie

1. Otwórz dane titanic
2. Spróbuj znaleźć zależności między zmiennymi wykorzystując:
   - Tabelę krzyżową
   - Współczynnik korelacji
   - Wykres typu scatterplot
   - Inne typy wykresów
3. Jaki % pasażerów każdej z klas przeżył katastrofę?
4. Kto miał większe szanse na przeżycie – kobiety czy mężczyźni?
5. Dla każej ze zmiennej przygotuj podstawowe statystyki opisowe