

Wydział Matematyki i Nauk Informacyjnych  
Politechnika Warszawska

Techniki wizualizacji danych  
semestr 21Z

Praca domowa nr. 3

Hubert Kozubek

Warszawa, 2021

# Spis treści

1.	Wprowadzenie . . . . .	1
2.	Dane do eksperymentu . . . . .	1
3.	Wizualizacje . . . . .	1
4.	Ankieta . . . . .	2
5.	Wnioski . . . . .	3

## 1. Wprowadzenie

Celem niniejszej pracy domowej było sprawdzenie czy określone problemy z czytaniem danych wciąż występują (problemy = iluzje na wykresach, błędy w odczytywaniu wykresu kołowego, błędy w pokazywaniu danych za pomocą wykresów 3D, problemy z tytułami na wykresie) oraz czy "dobre praktyki" opracowane lata temu (słupki lepsze niż kąty, kolor wykorzystywany z umiarem, unikanie ozdóbek), faktycznie działa w społeczeństwie które wykresami karmione jest od urodzenia.

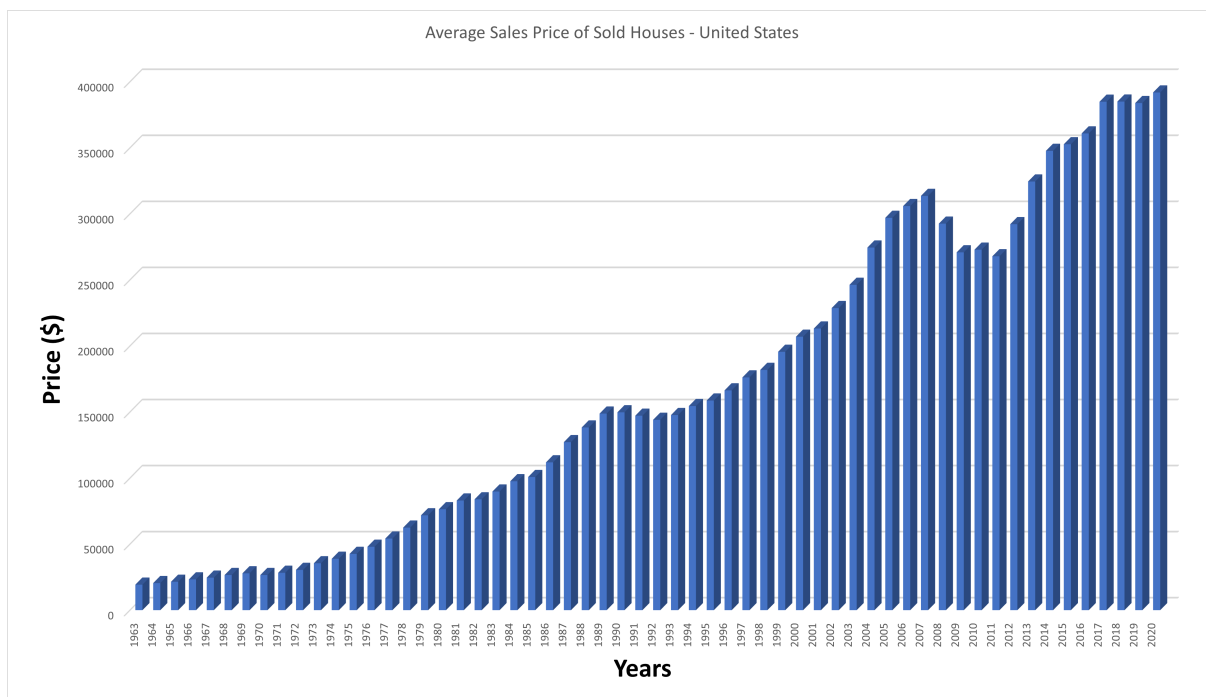
## 2. Dane do eksperymentu

Pomimo że wybór danych nie był kluczowy w przypadku tego projektu, starano się wybrać temat w miarę neutralny, aby ewentualne przekonania ankietowanych nie miały wpływu na głosowanie. Dodatkowo wykresy były wykonane na podstawie tych samych danych, jedyne co się zmieniało to sama wizualizacja danych.

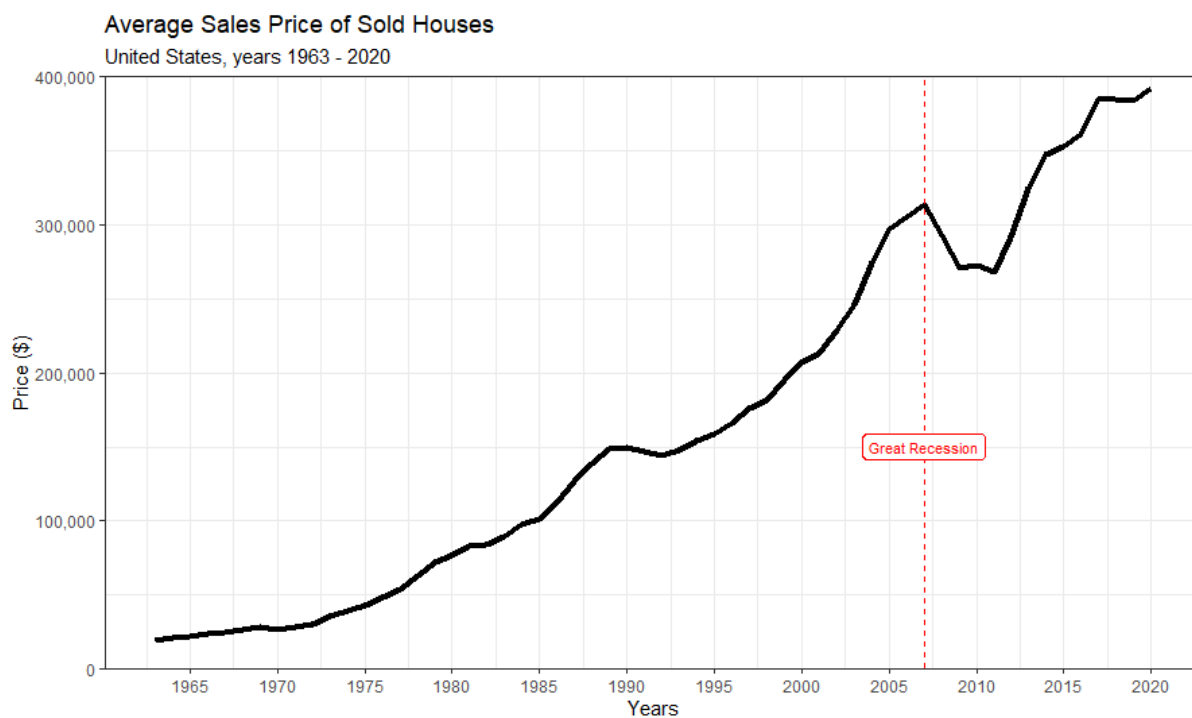
Jako motyw przewodni wizualizacji został wybrany temat średnich cen domów sprzedanych w USA, w latach 1963 - 2020. Dane wykorzystane dalej do wizualizacji zostały pobrane ze strony [www.data.gov](http://www.data.gov), a dokładnie z [catalog.data.gov/dataset/construction-price-indexes](https://catalog.data.gov/dataset/construction-price-indexes).

## 3. Wizualizacje

Analizę przeprowadzono tylko na części danych przedstawiające średnie ceny domów sprzedanych w danym roku. Do stworzenia wykresów użyto programu Excel oraz języka R wraz z pakietem ggplot2. Ostateczne wykresy, użyte później w ankiecie, przedstawiono na rys. 1 i rys. 2.



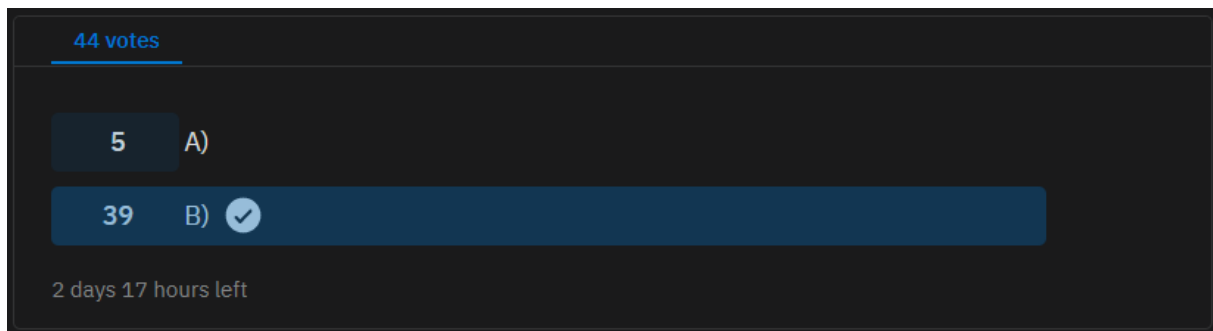
Rys. 1. Średnie ceny domów sprzedanych w USA, wykres stworzony w Excel'u



Rys. 2. Średnie ceny domów sprzedanych w USA, wykres stworzony przy pomocy pakietu ggplot2

#### 4. Ankieta

Ankieta której celem było zebranie opini na temat wykresów, została przeprowadzona na platformie reddit, na grupie poświęconej analizie danych ([Link do ankiety](#)). W czasie pisania tego sprawozdania odpowiedzi udzieliło 44 członków. Wyniki ankiety przedstawiono na rys. 3.



Rys. 3. Wyniki ankiety przeprowadzonej na członkach r/dataanalysis

## 5. Wnioski

Z wyników eksperymentu można jasno wywnioskować, że faworytem ankietowanych był wykres 2. Wynika to najprawdopodobniej z tego, że jest on bardziej czytelny ze względu na kontrastujące kolory oraz dobre rozmieszczenie podpisów na osiach. Wykres 1 natomiast jest trójwymiarowy, przez co zaburza percepcje osi y, dodatkowo aby odczytać wartość słupka trzeba się trochę namęczyć. Również podpisy na osi x są za bardzo stłoczone.

Warto wspomnieć że ankietę została przeprowadzona w środowisku osób, które interesują się analizą danych oraz jej wizualizacją, przez co nie reprezentują oni pełnego przekroju społeczeństwa. Dodatkowo zbyt mała wielkość próbki uniemożliwia uogólnienie wyników. Niemniej jednak nie oznacza to, że badanie to było bezużyteczne. Wiemy że zdecydowana większość grupy ponad 40 osób opowiadała się za wykresem nr 2, czyli za "dobrymi praktykami".