# PDRPy 2021/2022
# Praca domowa nr 2. Przetwarzanie danych i ich wizualizacja

Michał Szeląg Jan Szablanowski

Czerwiec 2022

# Temat prezentacji

Prezentacja przedstawia analizę zanonimizowanych danych z serwisów Stack Exchange. Każdy taki serwis to forum tematyczne, w którym użytkownicy mogą dodawać posty, komentować, głosować na najlepsze odpowiedzi i zdobywać odznaki.

# Zagadnienia badawcze

W pracy zbadano następujące zagadnienia:

- popularność warzenia piwa na tle innych alkoholi na forum ***Homebrewing***

# Zagadnienia badawcze

W pracy zbadano następujące zagadnienia:

- popularność warzenia piwa na tle innych alkoholi na forum ***Homebrewing***
- wpływ pandemii COVID-19 na podróżowanie na przykładzie forum podróżniczego ***Travel***

# Zagadnienia badawcze

W pracy zbadano następujące zagadnienia:

- popularność warzenia piwa na tle innych alkoholi na forum ***Homebrewing***
- wpływ pandemii COVID-19 na podróżowanie na przykładzie forum podróżniczego ***Travel***
- porównanie użytkowników Androida i iOS-a na podstawie forów ***Android*** i ***Apple***

# Struktura danych

Dane każdego serwisu mają taką samą strukturę:

- *Posts* - informacje o użytkownikach
- *PostHistory* - historia edycji posta
- *PostLinks* - powiązane posty
- *Comments* - komentarze do postów
- *Tags* - tagi
- *Users* - użytkownicy danego forum
- *Badges* - odznaki
- *Votes* - głosy

# Narzędzia badawcze

Do opracowania danych wykorzystano język Python wraz z bibliotekami:

- *pandas* - ramki danych
- *numpy* - operacje na wektorach i macierzach
- *matplotlib* - wykresy
- *scipy* - wybrane algorytmy np. interpolacja

# Analizu serwisu Homebrewing

# Charakterystyka serwisu

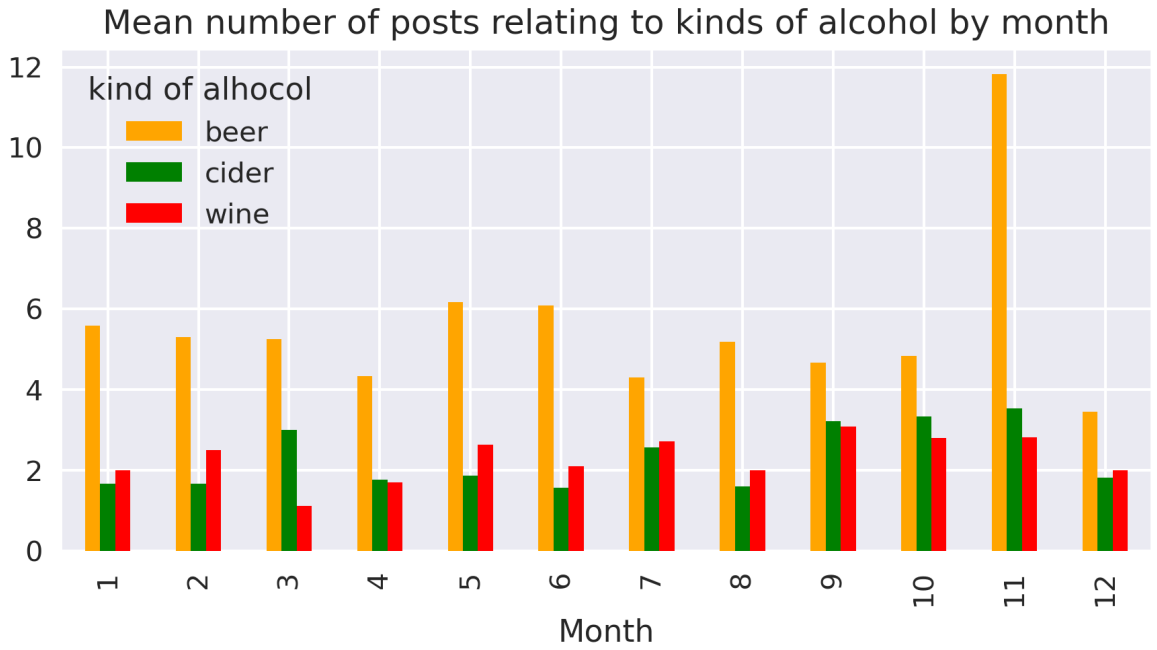
Serwis `www.homebrew.stackexchange.com` zajmuje się tematyką rzemieślniczej, domowej produkcji napojów alkoholowych. Użytkownicy na forum zadają pytania dotyczące fermentacji, zakażeń, dodatków czy innych technicznych szczegółów produkcji piwa, wina, cydru i innych wyrobów.

# Pytania badawcze

Podczas analizy danych serwisu zbadano następujące zagadnienia:

1. porównanie popularności pytań o piwo, wino i cydr
2. prześledzenie popularności forum w czasie
3. badanie najbardziej intensywnego okresu w życiu forum:
   - kiedy on nastąpił?
   - jaka część postów to pytania?
   - za jaką część postów odpowiadają najaktywniejsi użytkownicy?

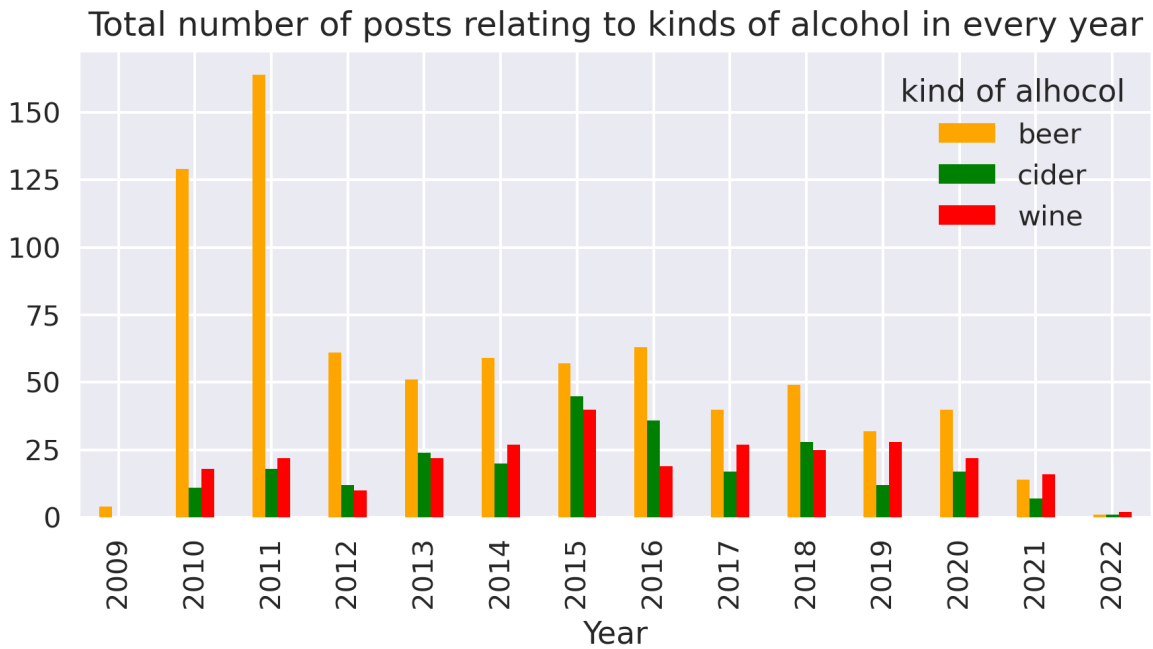
# Porównanie popularności pytań o piwo, wino i cydr




Średnia ilość postów dotyczących badanych rodzajów alkoholu na miesiąc ogółem

Zależność została zbadana poprzez wyliczenie średniej ilości postów zawierających odpowiednie tagi (*beer, wine, cider) na dany miesiąc w roku.*

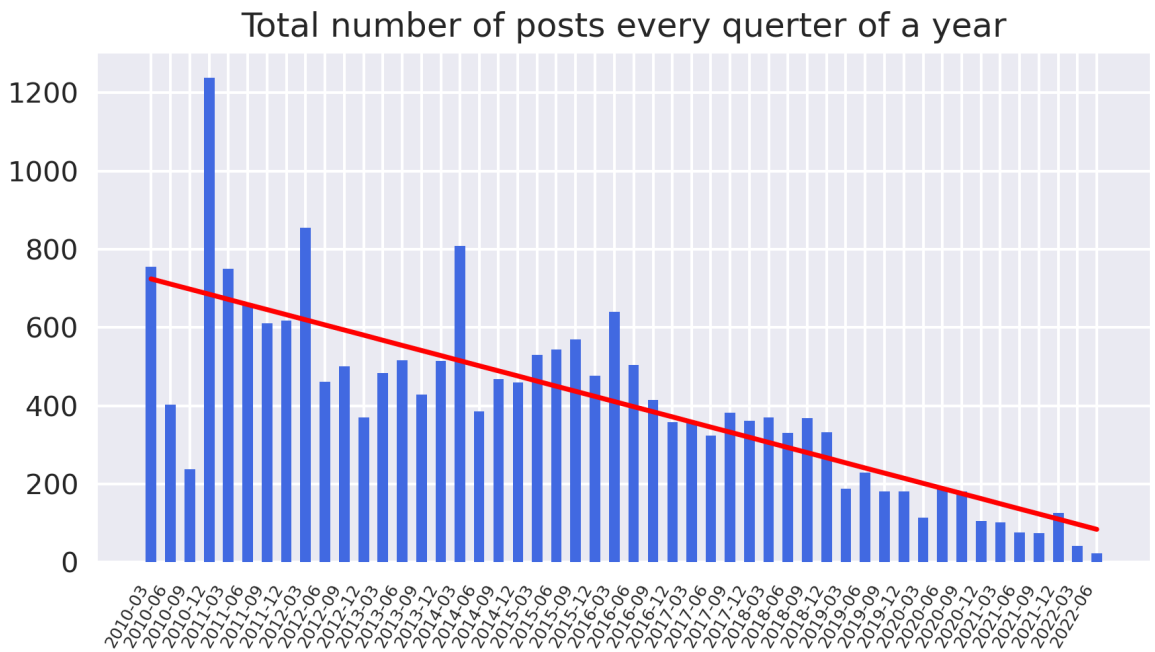
# Porównanie popularności pytań o piwo, wino i cydr




Ilość postów dotyczących badanych rodzajów alkoholu w danym roku

Zależność została zbadana poprzez wyliczenie ilości postów zawierających odpowiednie tagi (*beer, wine, cider) w danym roku.*

# Prześledzenie popularności forum w czasie



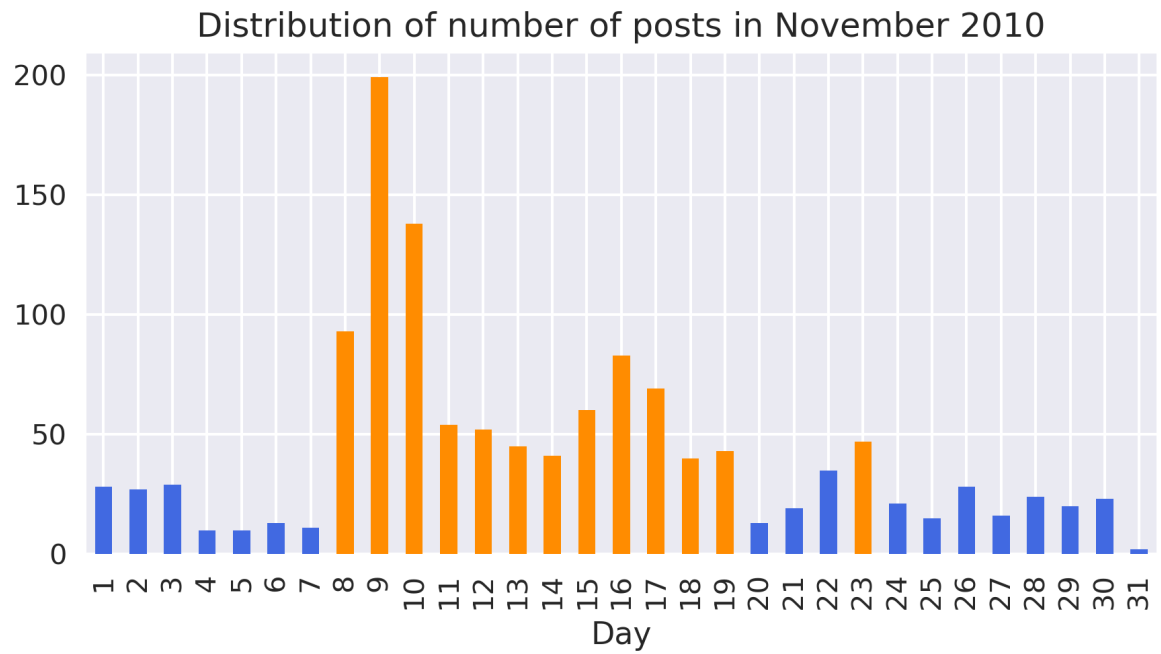

Ilość postów ogółem w danym kwartale

Zależność została zbadana poprzez wyliczenie ilości postów w danym kwartale. Dopasowana została do tego linia trendu.

# Wnioski

1. Pytania „piwne” pojawiają się dużo częściej niż te dotyczące cydru czy wina.
2. Znacząca średnia ilość pytań „piwnych” w listopadzie.
3. Znacząca ilość pytań „piwnych” w latach 2010 i 2011.
4. Znaczący wzrost postów ogółem w czwartym kwartale roku 2010 - prawdopodobnie odpowiada za skoki zaobserwowane wcześniej.
5. Ogólny spadek ilości nowych postów - wymieranie forum.

# Badanie najbardziej intensywnego okresu w życiu forum



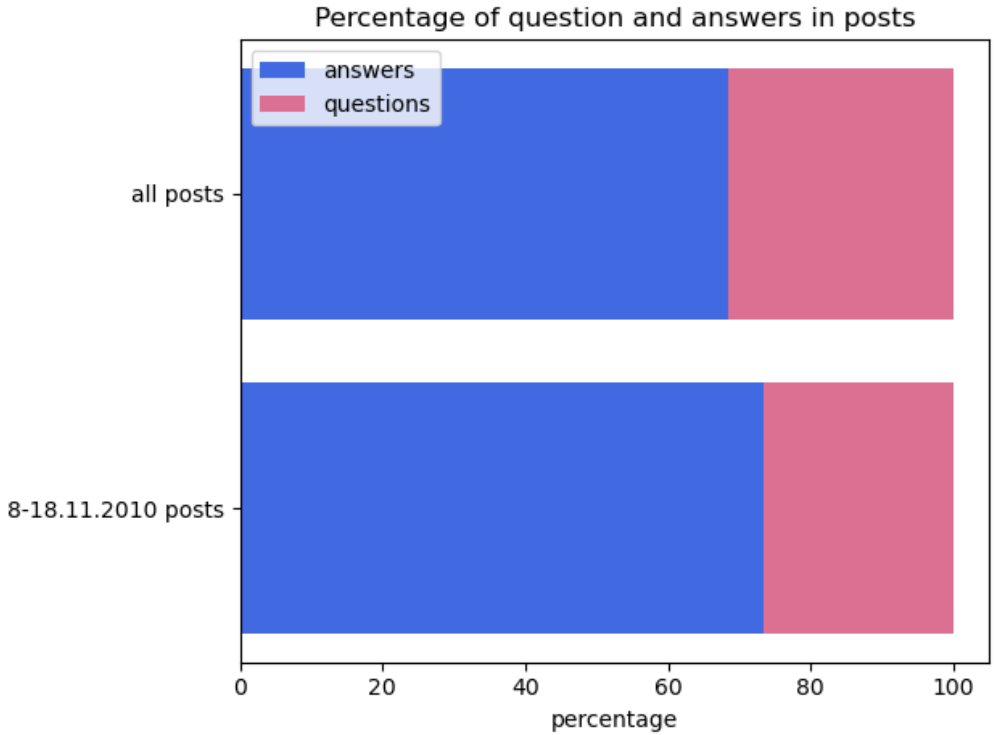

Rozkład ilości postów w listopadzie 2010

Wyliczenie ilości postów w danym dniu w tym okresie. Pomarańczowe słupki oznaczają przekroczenie 40-stu postów.

# Badanie najbardziej intensywnego okresu w życiu forum

Dalej badane będzie 10 dni począwszy od 8 listopada 2010 w porównaniu z danymi z całego życia forum. Porównanie między tymi zbiorami będą badane z uwagi na diametralną różnicę w ilości postów publikowanych przez dni następujące po 8 listopada 2010.

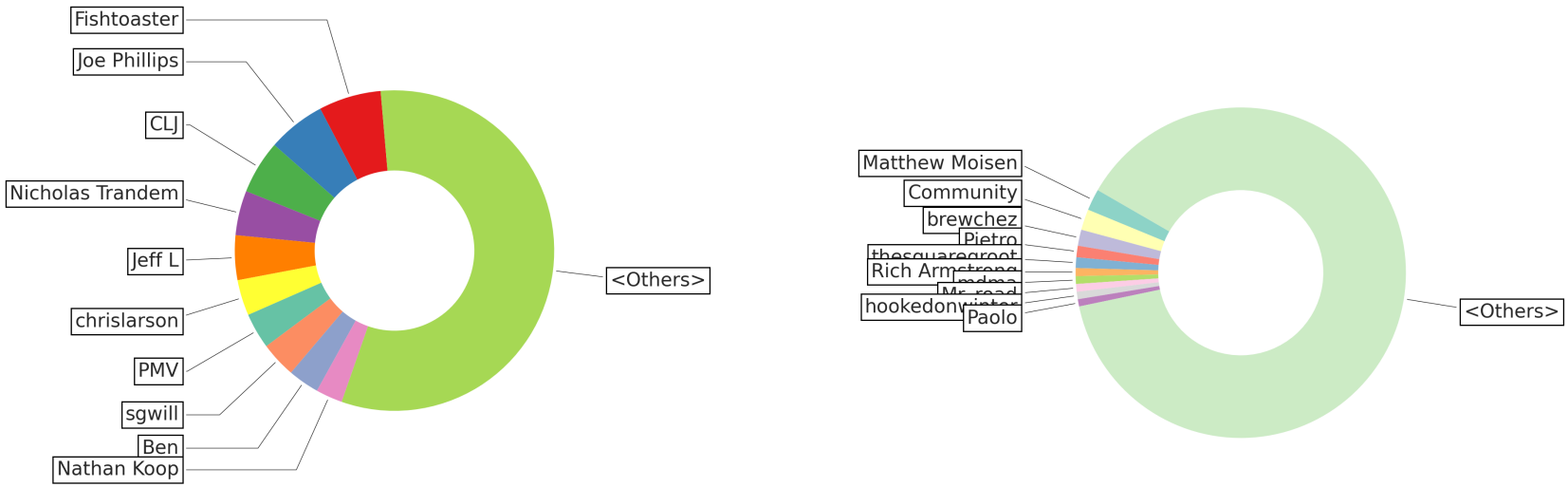
# Badanie najbardziej intensywnego okresu w życiu forum




Porównanie ilości pytań i odpowiedzi w postach (procentowo)

Zależność została zbadana grupując zbiory postów z danych okresów na pytania i odpowiedzi.

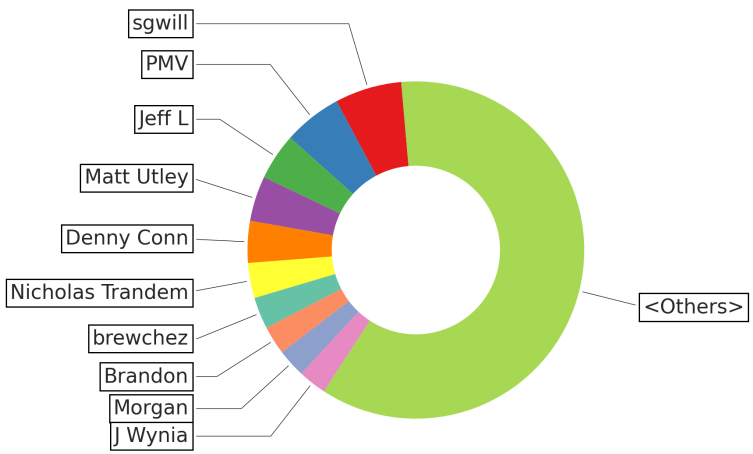
# Badanie najbardziej intensywnego okresu w życiu forum

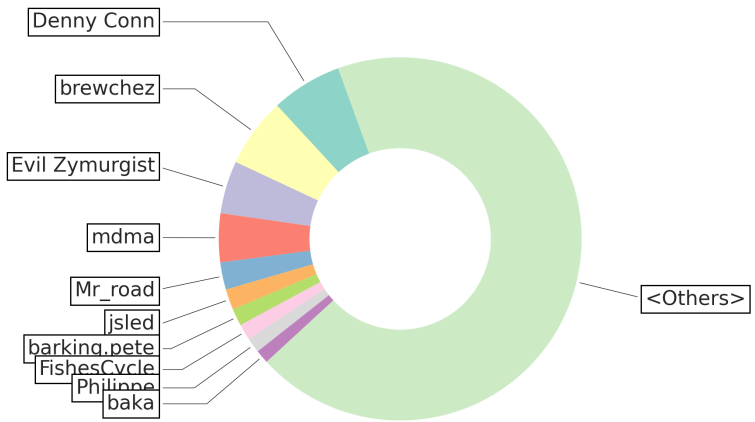



Porównanie jak dużą część pytań zadało top10 najbardziej aktywnych autorów pytań (po lewej okres 8-18.11.2010, po prawej całe forum).

Zależność została zbadana grupując posty według nazw autorów (z tabeli użytkowników).

# Badanie najbardziej intensywnego okresu w życiu forum




Porównanie jak dużą część odpowiedzi udzieliło top10 najbardziej aktywnych autorów odpowiedzi (po lewej okres 8-18.11.2010, po prawej całe forum).

Zależność została zbadana grupując posty według nazw autorów (z tabeli użytkowników).

# Wnioski końcowe

1. W intensywnym okresie nie zadano większej ilości pytań
2. Za to najbardziej aktywni autorzy pytań zadali ich w intensywnym okresie dużo więcej niż dzieje się to średnio.
3. Nie można stwierdzić podobnej zależności dla autorów odpowiedzi. Grono ekspertów przejawiało podobną aktywność w obu okresach.
4. Najaktywniejszy ogólnie ekspert jest czwartym najaktywniejszym ekspertem z intensywnego okresu.

# Analizu serwisu Travel

# Charakterystyka serwisu

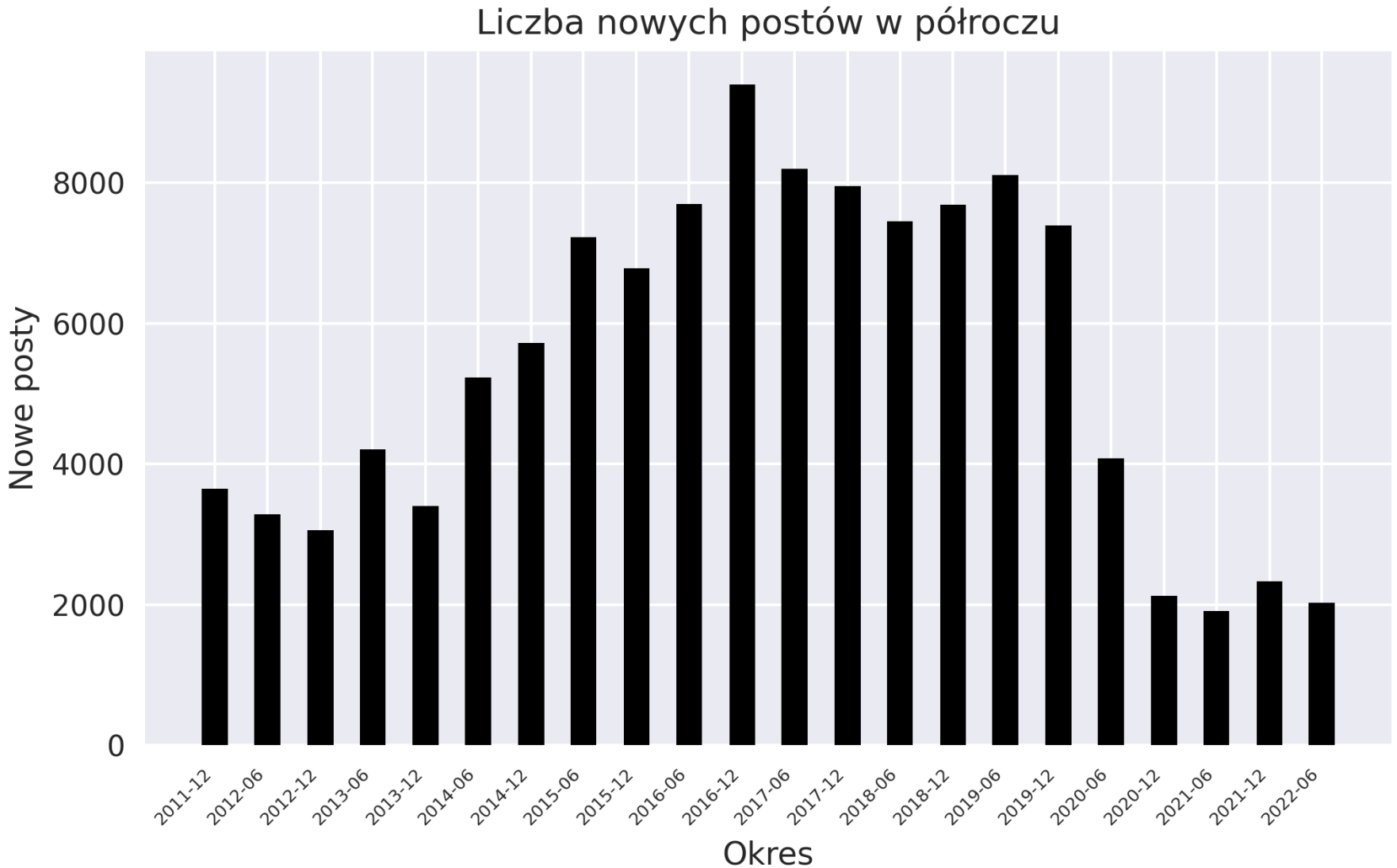
Tematyką serwisu `www.travel.stackexchange.com` jest podróżowanie. Użytkownicy wymieniają się na nim informacjami na temat ciekawych miejsc, sposobów transportu i noclegów.

# Pytania badawcze

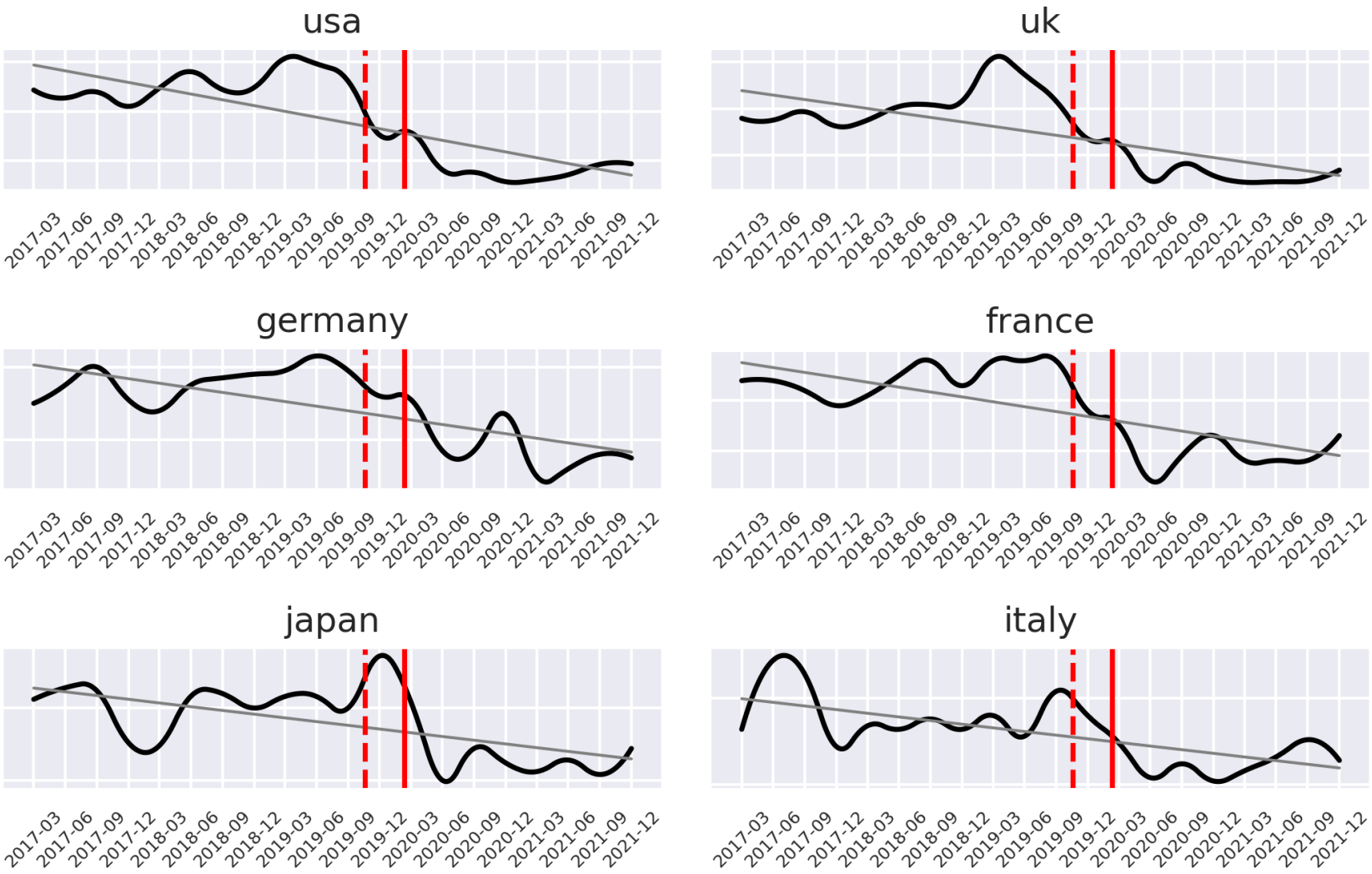
W ramach pracy badawczej postawiono następujące pytania:

1. jaki był wpływ pandemii COVID-19 na podróżowanie?
2. które państwa najmocniej odczuły skutki pandemii?
3. jak zmienił się poziom zaangażowania użytkowników forum?

# Wpływ pandemii na podróżowanie



# Wpływ pandemii na podróżowanie (c.d.)



# Wpływ pandemii na podróżowanie (c.d.)

Poniżej przedstawiono współczynniki kierunkowe linii trendu dla 6 najpopularniejszych krajów:

| **Kraj** | **wsp. kier.** |
|---|---|
| usa | -11.68 |
| uk | -9.56 |
| japan | -1.28 |
| germany | -1.25 |
| italy | -1.05 |
| france | -0.98 |

Metoda regresji liniowej

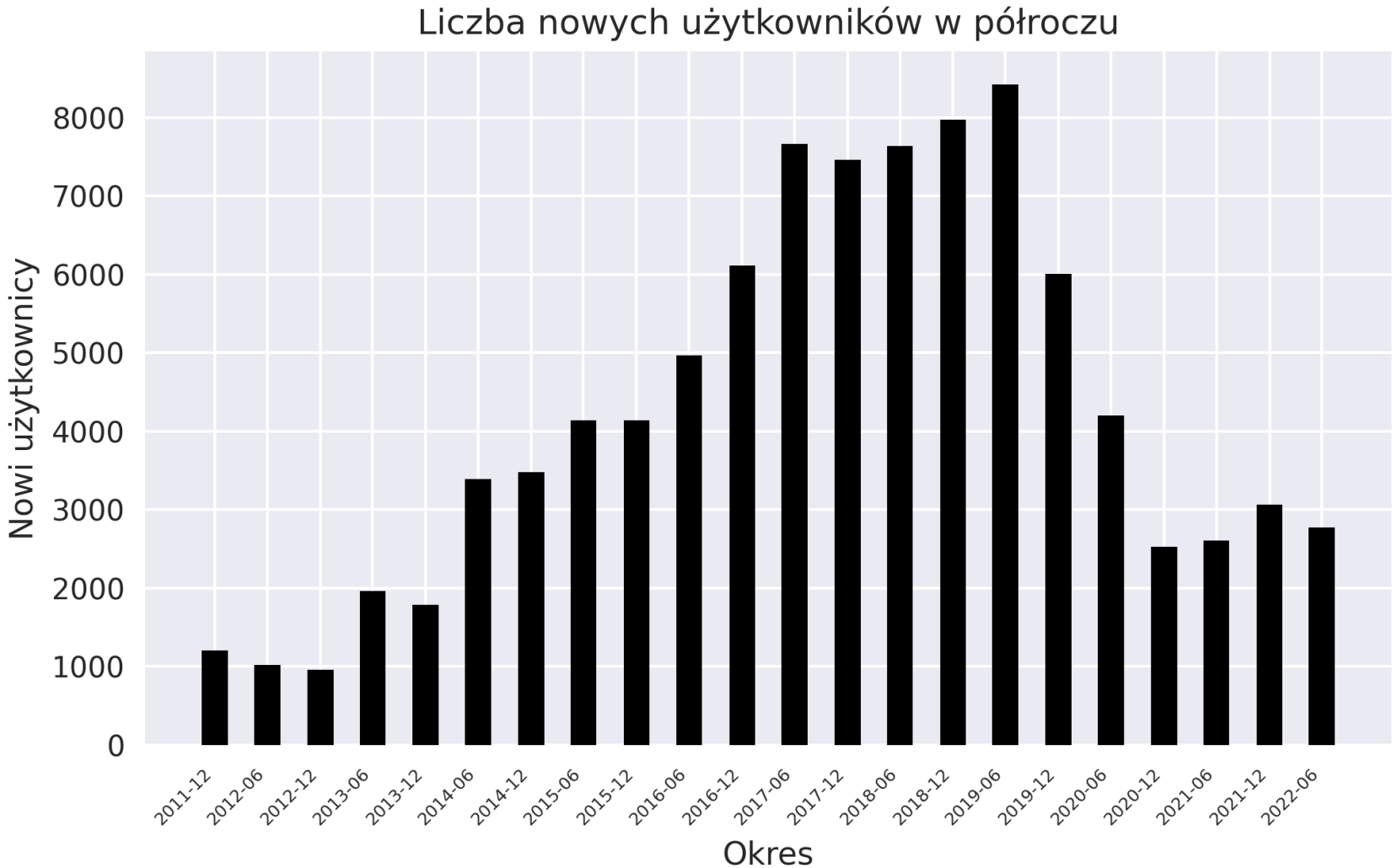
| **Kraj** | **wsp. kier.** |
|---|---|
| usa | -11.1 |
| uk | -8.69 |
| japan | -1.18 |
| germany | -1.16 |
| italy | -0.96 |
| france | -0.89 |

Mann-Kendall Trend Test

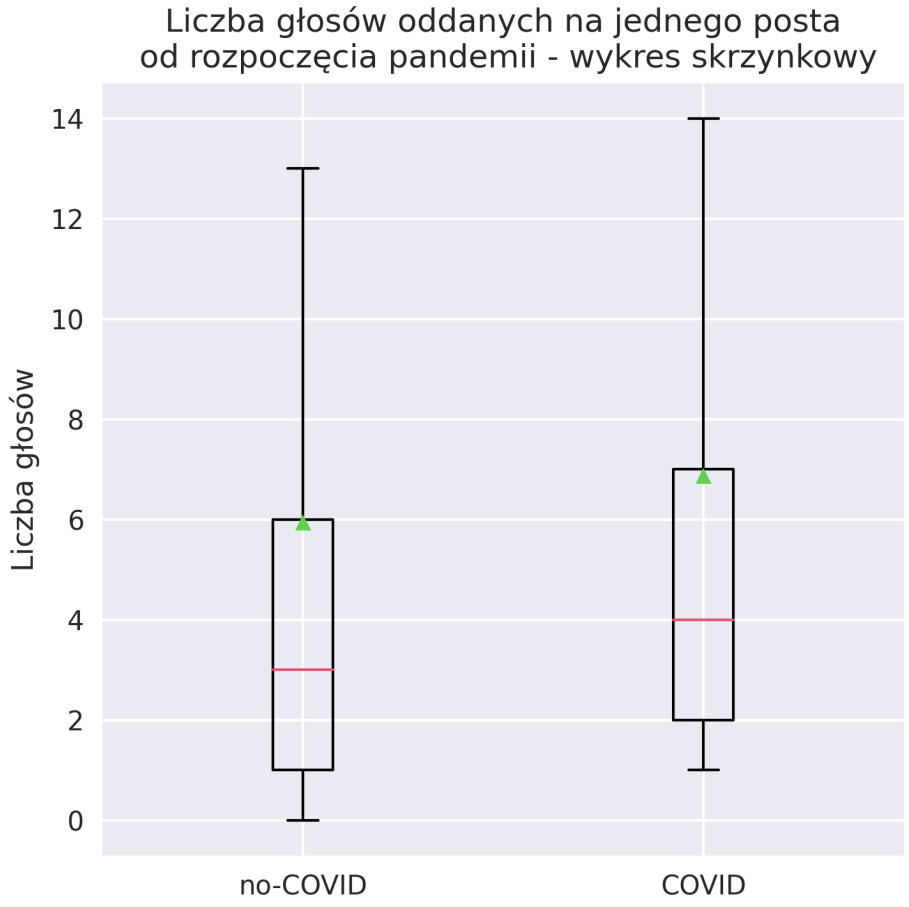
# Wpływ pandemii na podróżowanie - wnioski

1. z powodu pandemii największy spadek popularności na forum zanotowały USA i UK (może to wynikać z tego, że forum jest angielskojęzyczne)
2. u pozostałych państw stwierdzono podobne do siebie spadki
3. dla większości państw, w okresie między pierwszym zachorowaniem na COVID a ogłoszeniem pandemii liczba nowych postów była stała

# Nowi użytkownicy forum



# Zaangażowanie użytkowników




Posty zawierające wzmiankę o COVID cieszą się większym zainteresowaniem niż posty bez odniesień do pandemii.

# Porównanie serwisów Android i Apple (iOS)

# Charakterystyka serwisu

Serwis `www.android.stackexchange.com` jest poświęcony tematyce urządzeń z systemem Android. Zawiera pytania o funkcjonalności, rozwiązywanie problemów itp.

Serwis `www.apple.stackexchange.com` jest poświęcony podobnej tematyce, dotyczy jednak urządzeń firmy Apple. Żeby dane z obu serwisów były bardziej porównywalne, zawężono zbiór wątków z tego serwisu do tych, które dotyczą urządzeń z iOS.

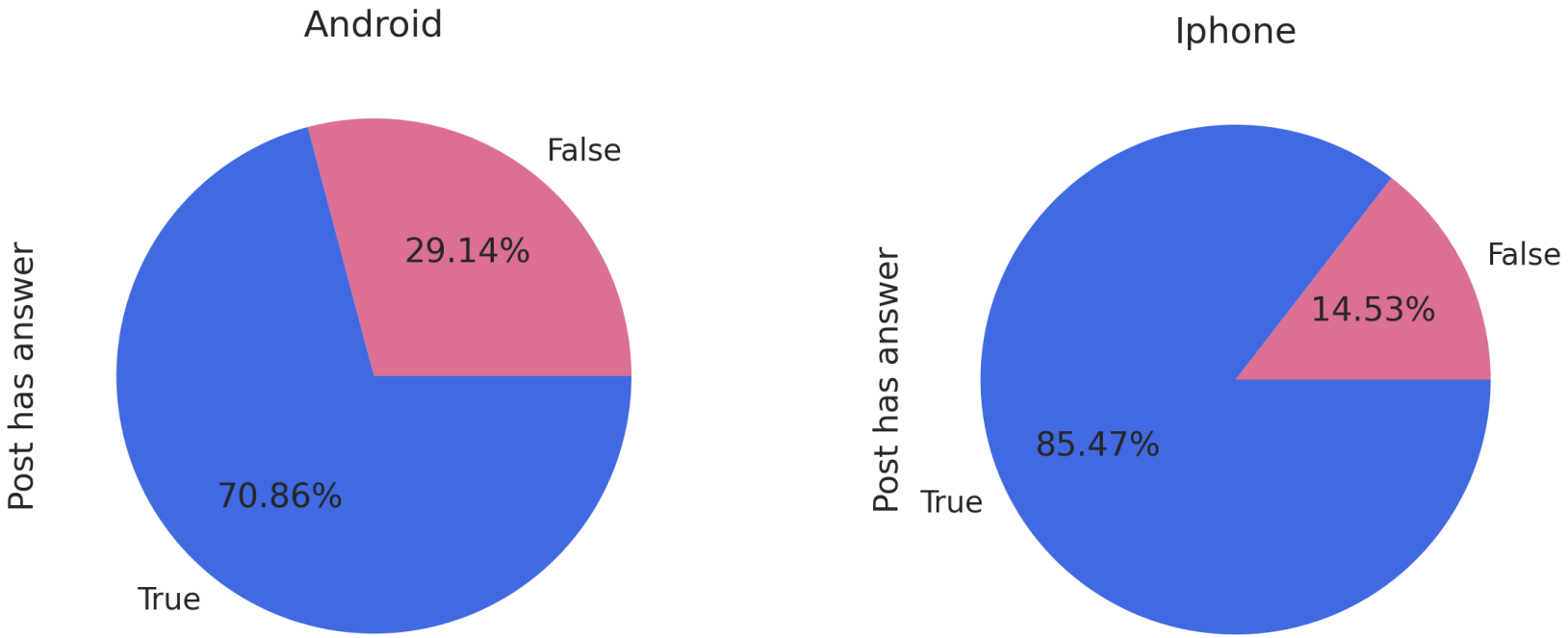
# Pytania badawcze

W ramach pracy badawczej porównano to, jak pomocni są użytkownicy forów Android i Apple w pytaniach dotyczących urządzeń mobilnych.

W szczególności porównywano następujące wielkości:

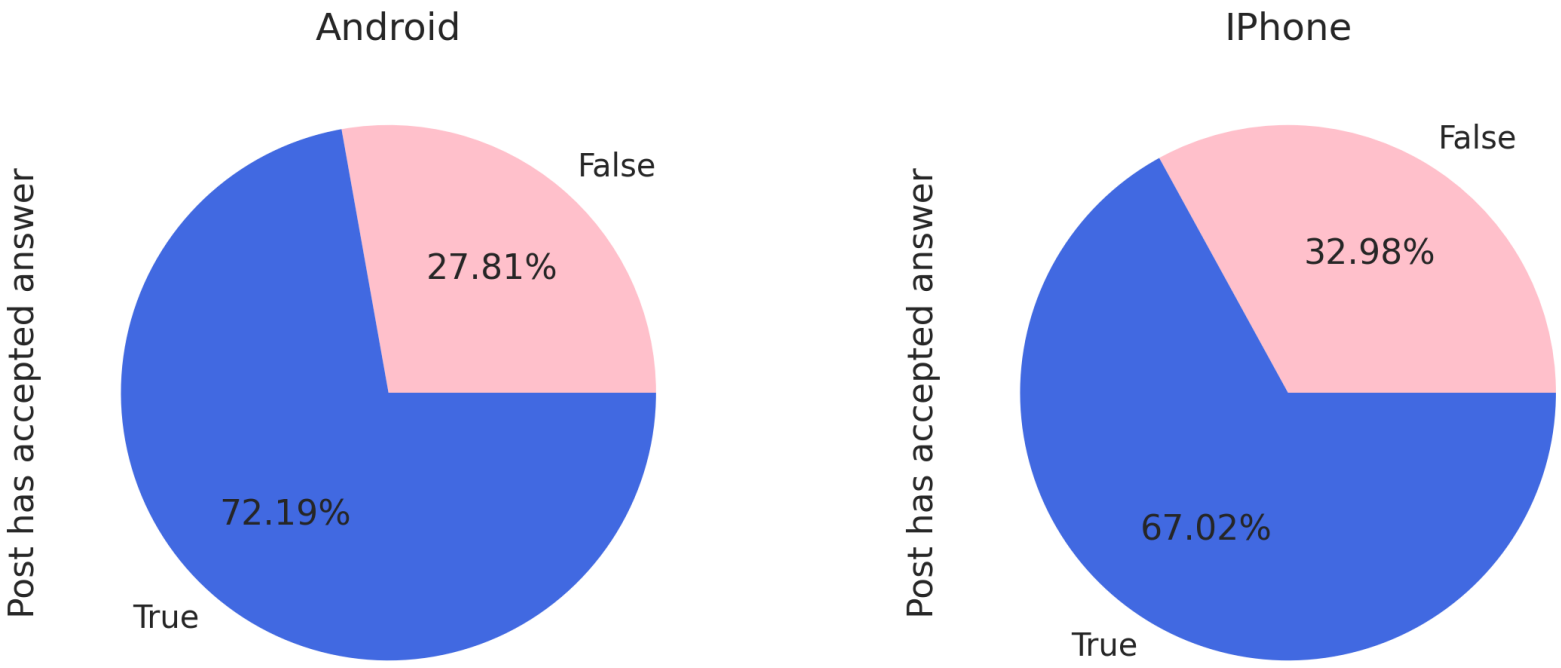
1. procent postów, które otrzymały jakąkolwiek odpowiedź
2. procent postów z odpowiedziami, z których jedna otrzymała akceptację

# Posty, które mają odpowiedź




Porównanie względnej liczby postów, które dostały odpowiedź

# Posty, które mają *zaakceptowaną* odpowiedź




Porównanie względnej liczby postów, które dostały *zaakceptowaną* odpowiedź

# Android vs iOS - wnioski

- użytkownicy iOS-a są bardziej pomocną społecznością (więcej postów z udzieloną odpowiedzią)
- może to wynikać z tego, że jest więcej użytkowników serwisu Apple'a (ale nie musi)
- natomiast posty na forum Androida mają względnie więcej zaakceptowanych odpowiedzi
- możliwe, że forum Androida posiada więcej ekspertów i odpowiedzi są trafniejsze

Dziękujemy za uwagę!