PDRP Praca domowa 2

Wojciech Klusek, Aleksander Kuś

Wykorzystane zbiory danych

Do przeprowadzenia analizy wykorzystaliśmy następujące zbiory danych z serwisu Stack Exchange:

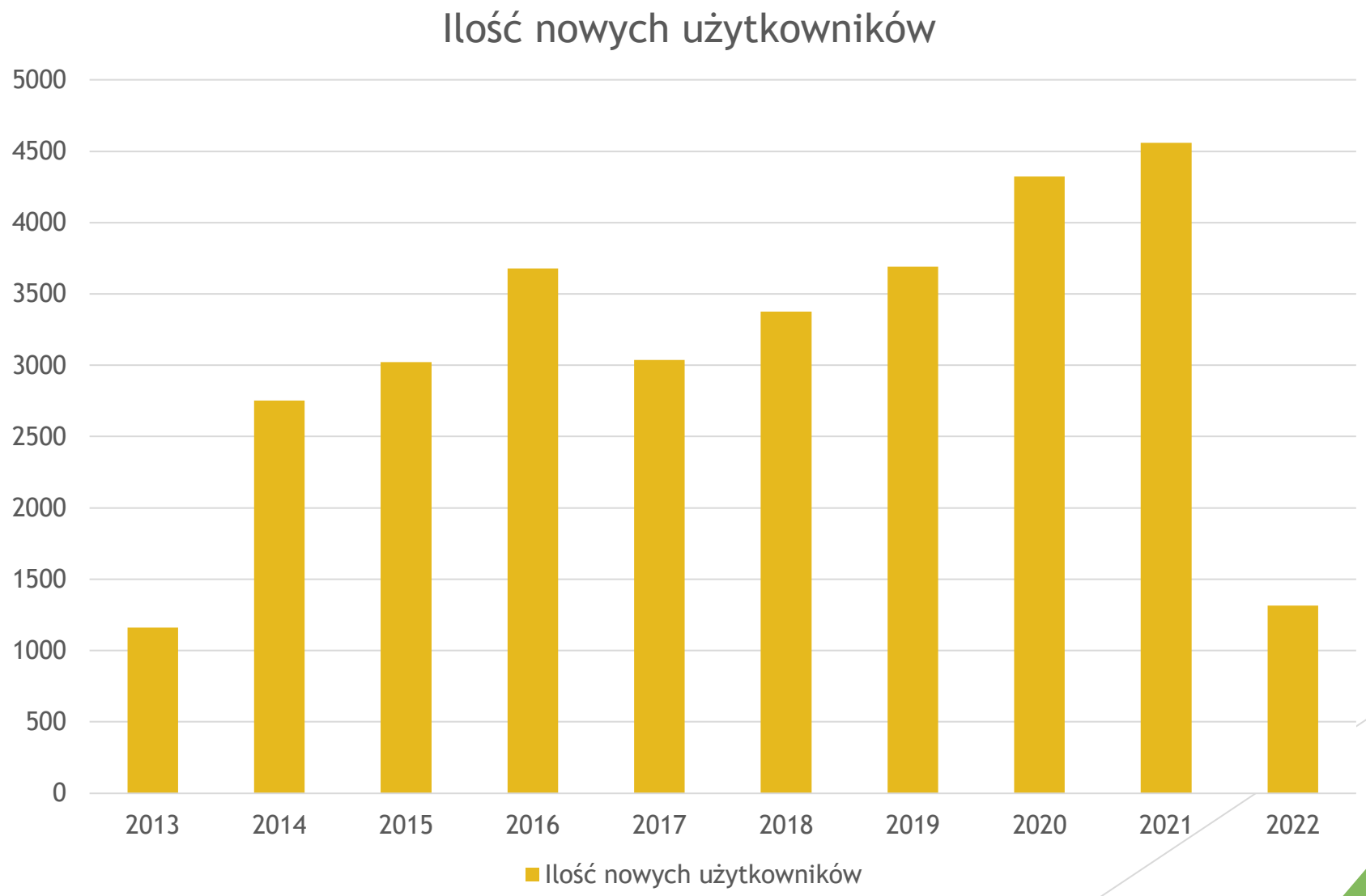
- Astronomy
- Devops
- Ebooks

Ilość nowych użytkowników w poszczególnych latach

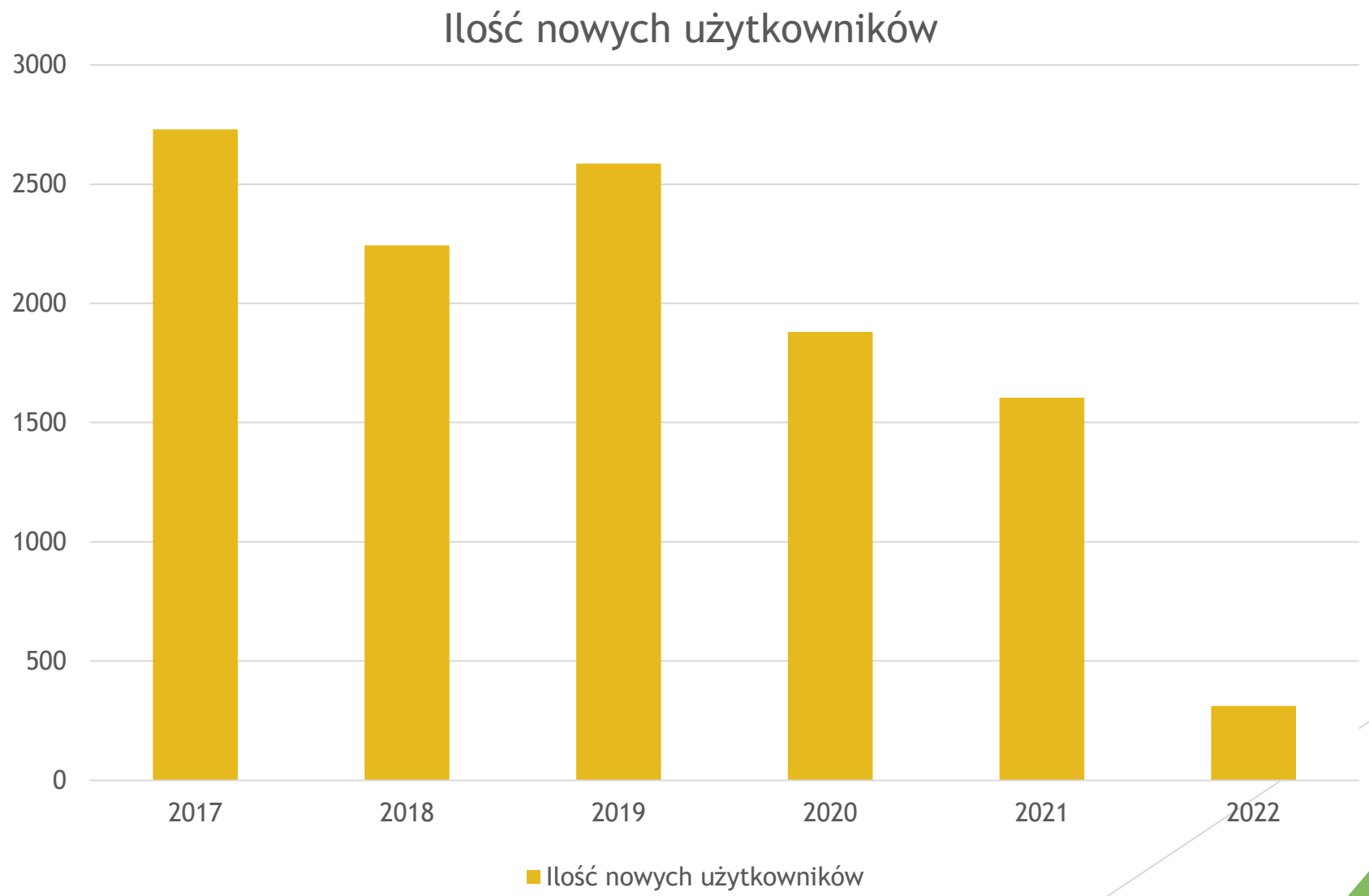
- Zapytanie to głównie miało na celu zbadać, jak w poszczególnych latach zmieniała się ilość nowych użytkowników każdego z 3 serwisów, które wybraliśmy.

```
getYearsWhenMostUsersSignedUp <- function(usersDataFrame)
{
  usersDataFrame %>%
    group_by(Year = format(as.Date(CreationDate), "%Y")) %>%
    summarize(Count = n()) %>%
    arrange(desc(Count)) %>%
    head(10)
}
```

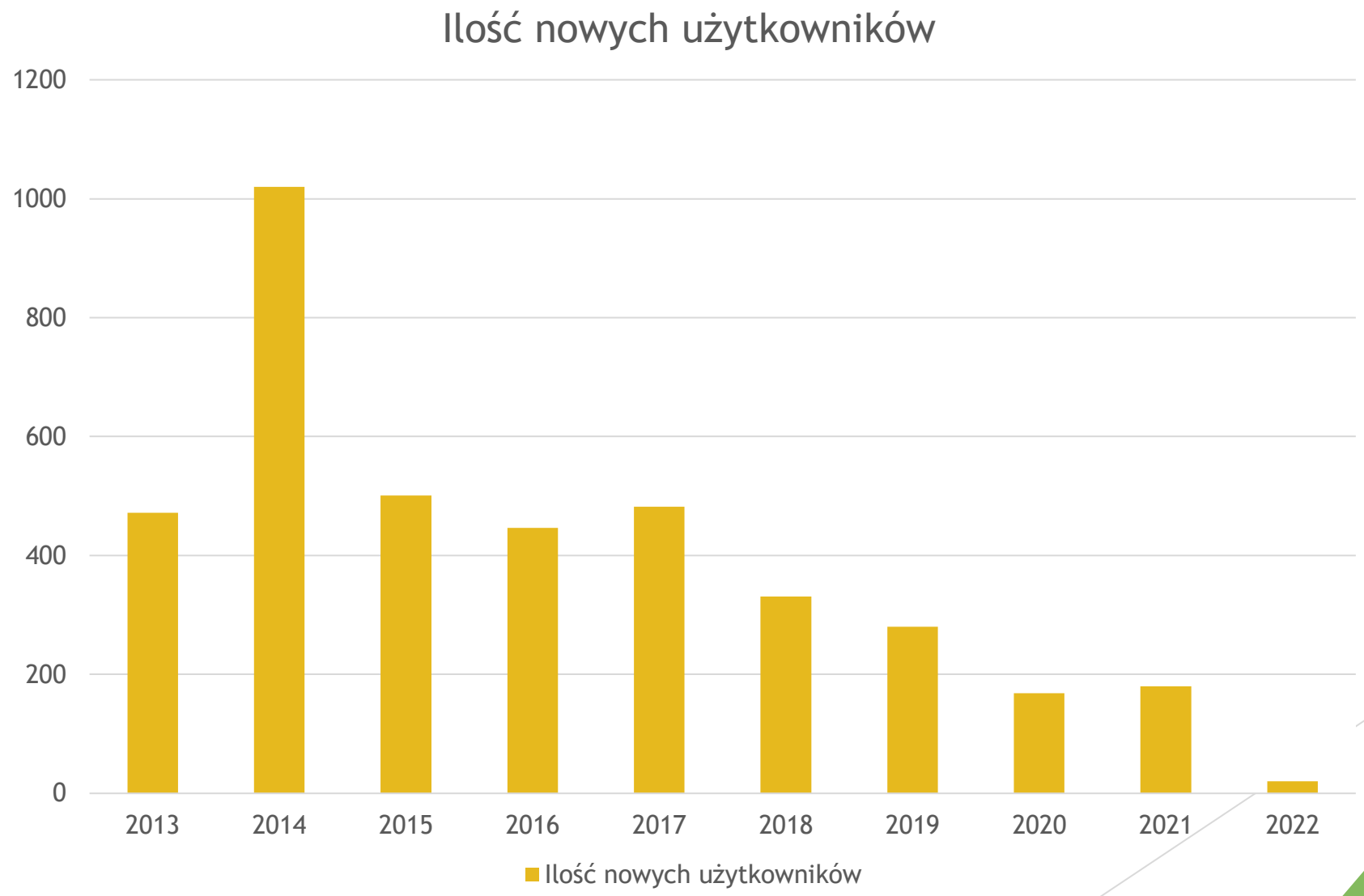
Serwis Astronomy



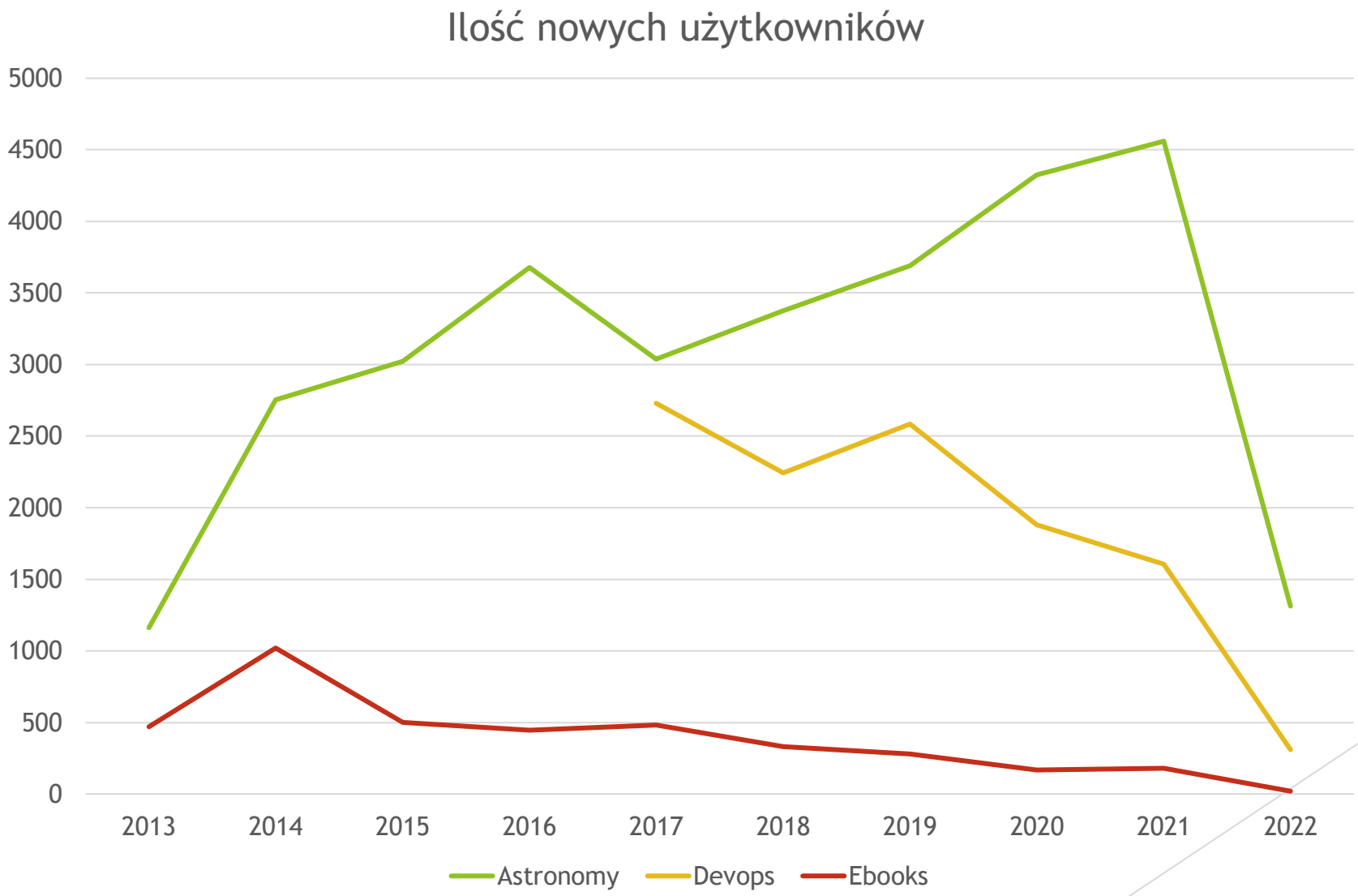
Serwis Devops



Serwis Ebooks



Wszystkie serwisy



10 lokalizacji z najlepszymi wynikami postów

- Zapytanie to głównie miało na celu zbadać jak w poszczególnych latach zmieniała się ilość nowych użytkowników każdego z 3 serwisów, które wybraliśmy.

```
getLocationsWithBestScore <- function(usersDataFrame, postsDataFrame)
{
  usersDataFrame %>%
    inner_join(postsDataFrame, by=c('Id' = 'OwnerUserId')) %>%
    filter(Location != '') %>%
    group_by(Location) %>%
    summarise(Scores = sum(as.numeric(Score))) %>%
    arrange(desc(Scores)) %>%
    head(10)
}
```

Astronomy

Lokalizacja	Sumaryczny wynik
United Kingdom	9279
Taipei, Earth	5016
A small planet somewhere in the vicinity of Betelgeuse	3130
Copenhagen, Denmark	2795
Houston, TX	2556
Virginia, USA	2197
California	1904
USA	1527
Munich, Germany	1416
Europe	1291

Devops

Lokalizacja	Sumaryczny wynik
Utrecht, Netherlands	1163
Brighton, England, United Kingdom	1000
Nantes, France	979
Ramat Gan, Israel	897
Germany	835
London, UK	718
www.abitmore-scm.com	626
Ottawa, ON, Canada	585
San Francisco Bay Area	556
Kansas City, KS, United States	374

Ebooks

Lokalizacja	Sumaryczny wynik
Germany	961
Pittsburgh, PA	587
Italy	565
Canada	439
Portland, OR	406
Houston, TX	353
New York, NY	270
Milan, Italy	221
Death Star	193
Springfield, IL, USA	154

Funkcja wykorzystana do operacji na tagach

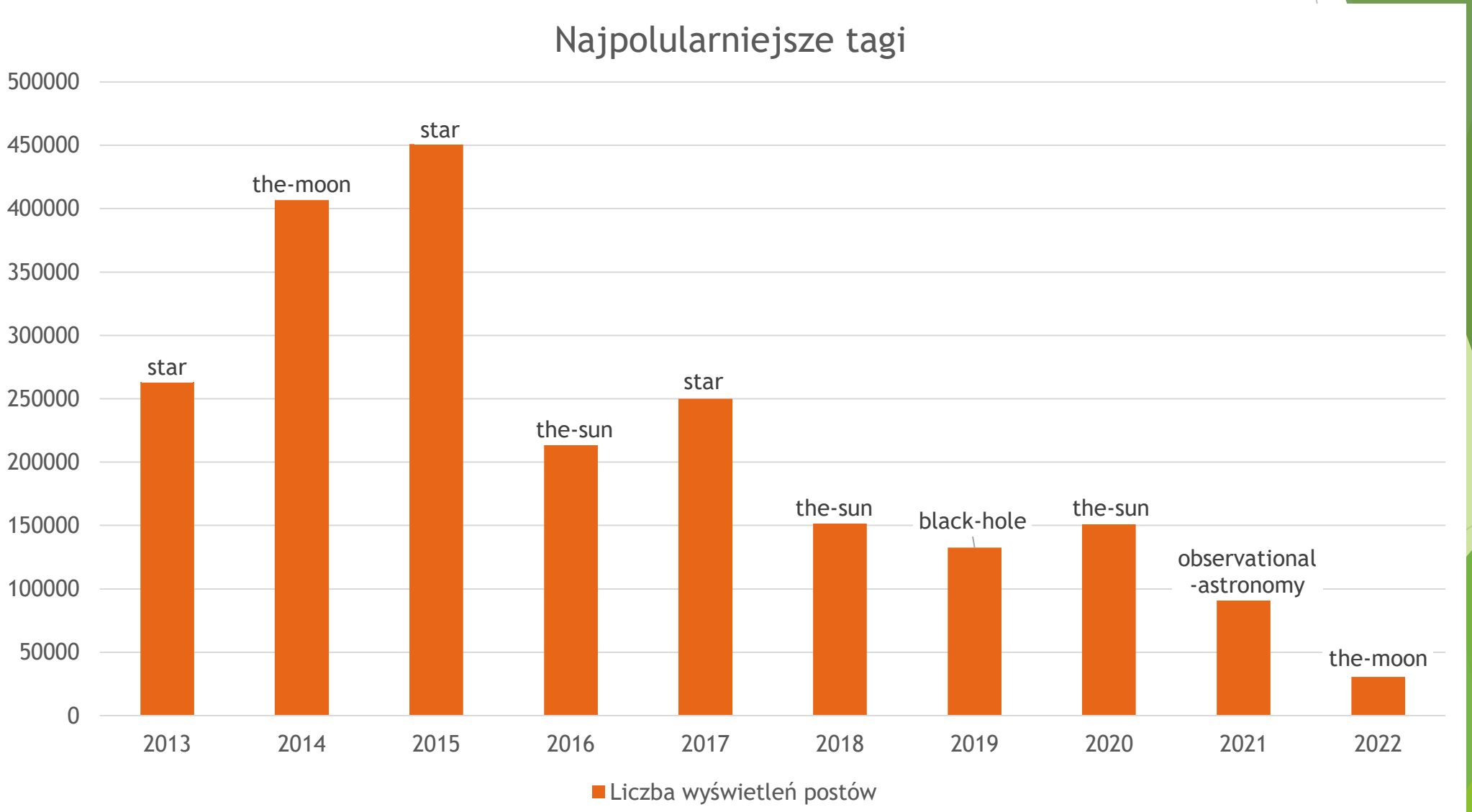
```
groupTags <- function(postsDataFrame)
{
  strings <- postsDataFrame[, c("Tags")]
  df <- data.frame(matrix(ncol = 2, nrow = 0))
  colnames(df) <- c("PostId", "Tag")
  for (i in 1:length(strings))
  {
    if (is.na(strings[i]))
    {
      next
    }
    splited <- strsplit(gsub('<', '', strings[i]), split = ">")
    for (j in 1:length(splited[[1]]))
    {
      df[nrow(df) + 1,] = c(postsDataFrame[i, 1], splited[[1]][j])
    }
  }

  df
}
```

Najpopularniejsze tagi w każdym roku

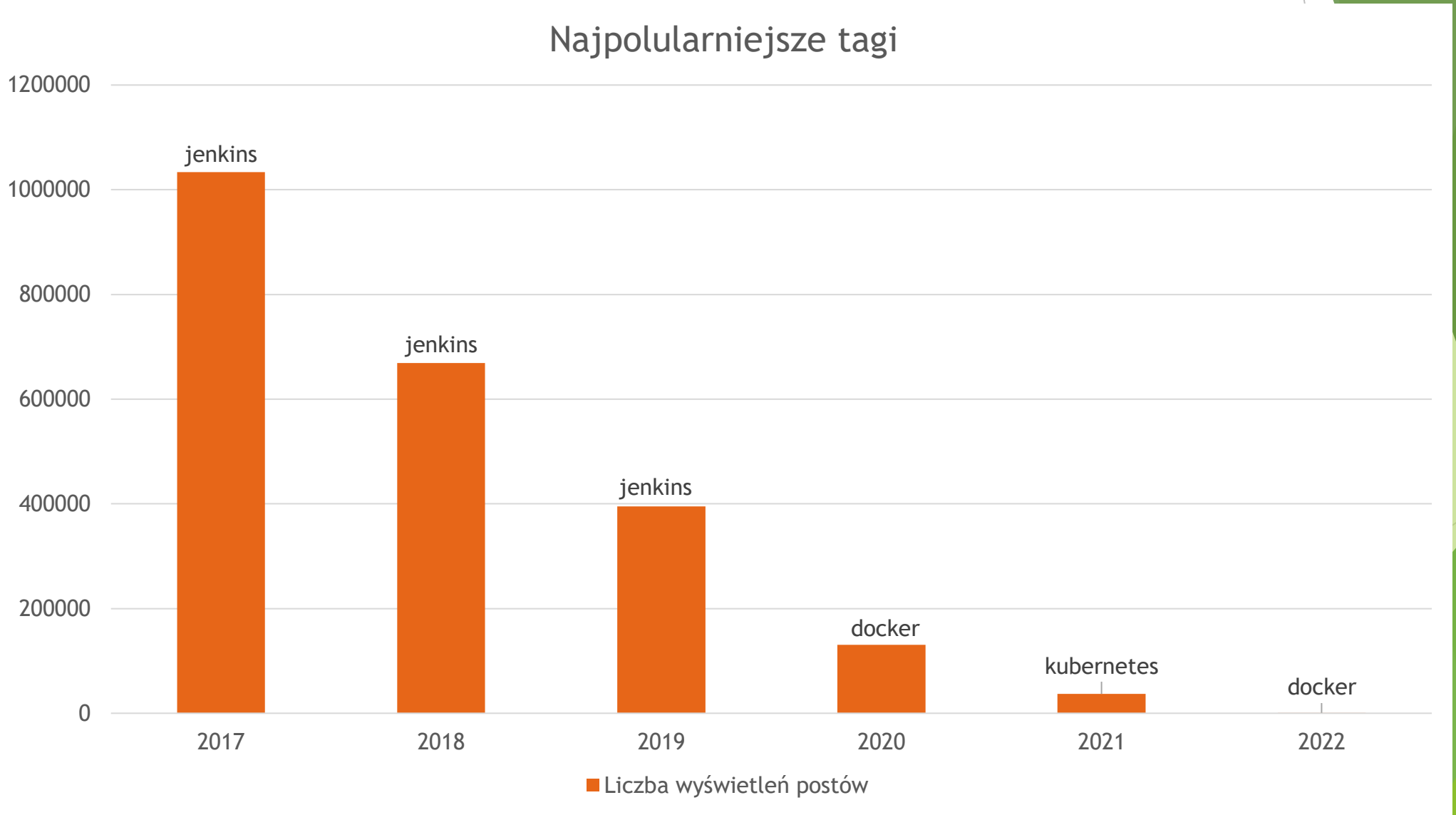
- Zapytanie to miało na celu zbadać jakie były najpopularniejsze tematy w poszczególnych latach.

```
mostPopularTagInYears <- function(postsDataFrame)
{
  groupedTags <- groupTags(postsDataFrame)

  groupedTags %>%
    inner_join(postsDataFrame, c('PostId' = 'Id')) %>%
    group_by(Year = format(as.Date(CreationDate), "%Y"), Tag) %>%
    summarise(Views = sum(as.numeric(ViewCount))) %>%
    arrange(desc(Views)) %>%
    select(Tag, Year, Views) -> groupedTagsByYear

  groupedTagsByYear %>%
    group_by(Year) %>%
    summarise(Tag = Tag[which.max(Views)], Views = Views[which.max(Views)]) %>%
    arrange(desc(Views))
}
```

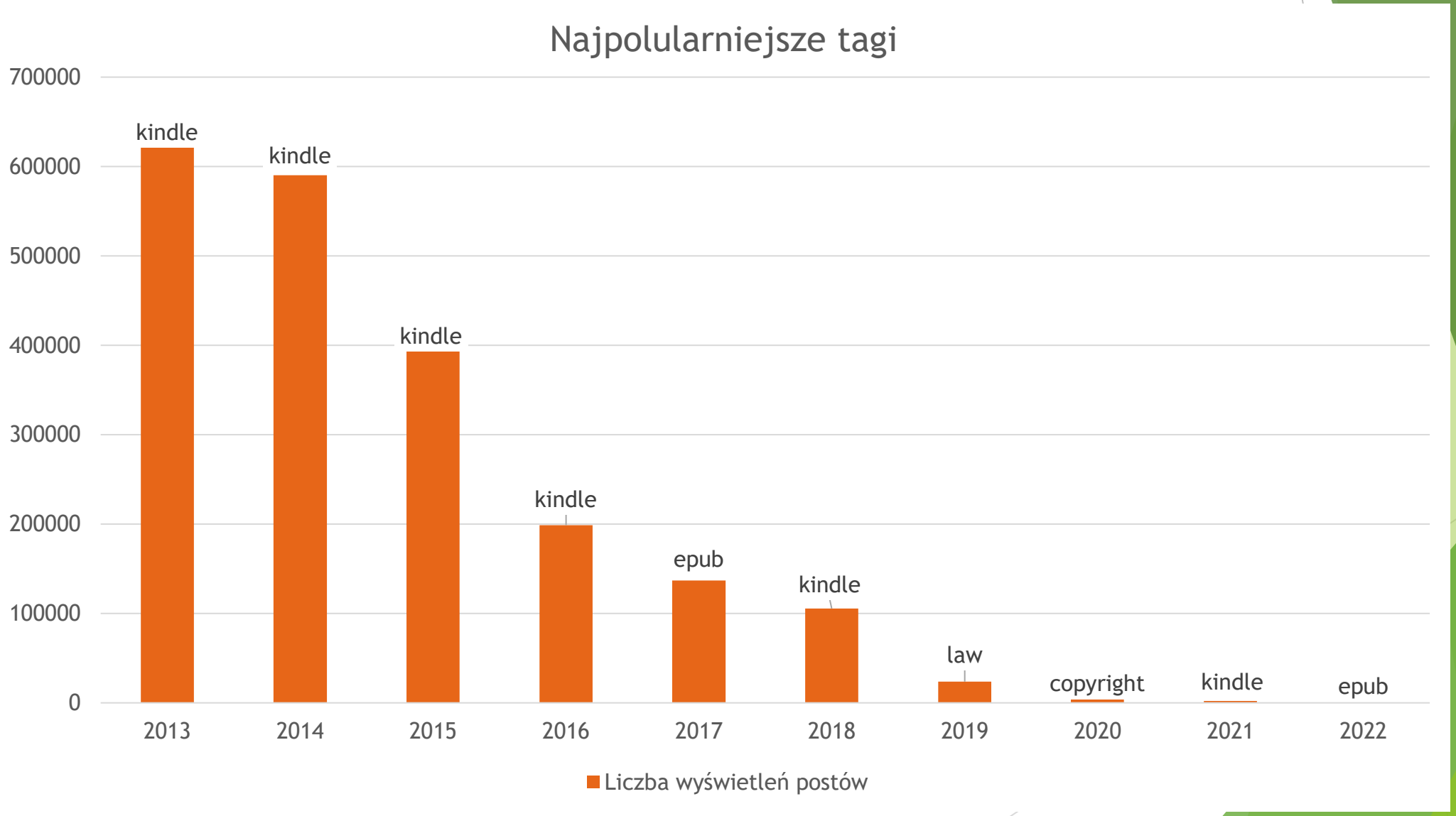
Astronomy


Najpolularniejsze tagi
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
star
the-moon
star
the-sun
star
the-sun
black-hole
the-sun
observational
-astronomy
the-moon
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Liczba wyświetleń postów

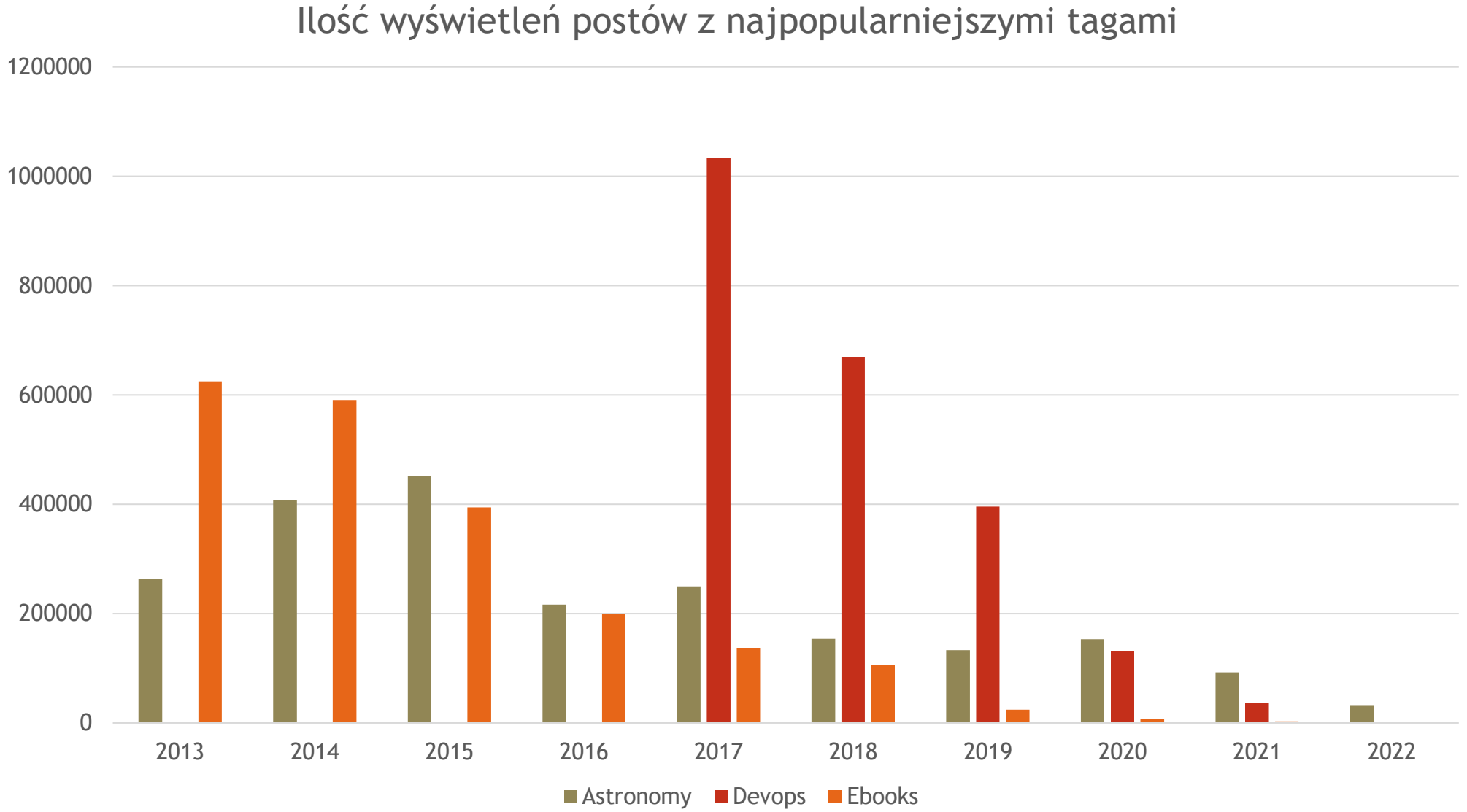
Devops


Najpolularniejsze tagi
1200000
1000000
800000
600000
400000
200000
0
jenkins
jenkins
jenkins
docker
kubernetes
docker
2017
2018
2019
2020
2021
2022
Liczba wyświetleń postów

Ebooks


Najpolularniejsze tagi
700000
600000
500000
400000
300000
200000
100000
0
kindle
kindle
kindle
kindle
epub
kindle
law
copyright
kindle
epub
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Liczba wyświetleń postów

Wszystkie serwisy



Tagi z najlepszym średnim wynikiem

- Zapytanie to miało na celu zbadać jakie były tagi, którymi otagowane posty średnio otrzymywały najlepsze oceny.

```
tagsWithBestAvgScore <- function(postsDataFrame)
{
  groupedTags <- groupTags(postsDataFrame)

  groupedTags %>%
    inner_join(postsDataFrame, c('PostId' = 'Id')) %>%
    group_by(Tag) %>%
    summarise(AvgScore = mean(as.numeric(Score)), Count = n()) %>%
    arrange(desc(AvgScore)) %>%
    filter(Count >= 10) %>%
    head(10)
}
```

Astronomy

Tag	Średni wynik	Liczba postów
io	12,7	14
9th-planet	10,5	34
trappist-1	9,09	23
naked-eye	8,71	56
event-horizon-telescope	8,59	29
solar-system-evolution	8,39	23
plasma-physics	8,19	31
oort-cloud	8,14	36
explosion	8,11	18
spiral-arms	8	12

Devops

Tag	Średni wynik	Liczba postów
terminology	12,7	85
ansible-vault	9,64	11
artifacts	9,4	30
immutable-servers	8,73	11
culture	8,71	87
process	8,67	18
project-management	8,06	17
sre	7,82	22
capacity-planning	7,7	10
backup	7,06	16

Ebooks

Tag	Średni wynik	Liczba postów
wifi	12,9	10
free	9,82	11
e-ink	8,73	22
law	7,73	11
linux	7,59	17
software	7,56	27
drm	7,55	71
collections	7,3	10
metadata	6,63	27
publishing	6,62	56

Dziękujemy za uwagę